Data Analysis Problems in Single-Case Evaluation: Issues for Research on Social Work Practice

Allen Rubin and Karen S. Knox

Research on Social Work Practice 1996; 6; 40
DOI: 10.1177/104973159600600103

Data Analysis Problems in Single-Case Evaluation: Issues for Research on Social Work Practice

Allen Rubin
University of Texas at Austin
Karen S. Knox
Southwest Texas State University

Data analysis problems, particularly involving the likelihood of obtaining visually ambiguous graphs, pose a barrier to efforts to promote increased use of single-case evaluation by practitioners. The debate about whether these efforts will eventually achieve sufficient success has thus far paid little attention to the impact of data analysis problems on practitioners' propensity to sustain a commitment to conducting single-case evaluations. This article uses findings from an evaluation of a cognitive-behavioral intervention with adolescent sex offenders to illustrate these data analysis problems and develop issues for research on social work practice and the teaching of single-case evaluation.

In 1984, in the aftermath of early enthusiasm about the promise of single-case evaluation for integrating research and practice, the Council on Social Work Education (CSWE) adopted new accreditation standards implying the need for all schools of social work to prepare students to evaluate their own practice. Although the standards did not explicitly specify the use of single-case evaluation techniques as the only way for practitioners to evaluate their own practice, the emergence of the notion of empirically evaluating one's own practice coincided with the advent of single-case evaluation, and the two concepts are generally associated with one another (Bloom & Fischer, 1982; Jayaratne & Levy, 1979).

Throughout the 1980s, as more schools of social work began teaching single-case evaluation as a way to conform to the new accreditation standards, studies consistently reported disappointing findings about the extent to which

Authors' Note: Correspondence may be addressed to Allen Rubin, School of Social Work, University of Texas at Austin, Austin, TX 78712. The authors gratefully acknowledge the assistance of Kelly Larson in preparing this article.

Research on Social Work Practice, Vol. 6 No. 1, January 1996 40-65

practitioners who learn about single-case evaluation as students eventually use it in their professional practice. In light of these studies, social work educators currently disagree as to whether single-case evaluation has lived up to its promise (Blythe, 1983, 1990; Briar, 1990; Corcoran, 1990; Dean & Reinherz, 1986; Dolan & Vourlekis, 1983; Fortune, 1982; Gingerich, 1977, 1984, 1990; Ivanoff, Blythe, & Briar, 1987; Mutschler, 1984; Nelsen, 1981, 1990; Penka & Kirk, 1991; Reinherz, 1990; Richey, Blythe, & Berlin, 1987; Robinson, Bronson, & Blythe, 1988; Siegel, 1983; Simons, 1987; Stern, 1990; Thyer, 1990; Tolson, 1990; Welch, 1983). Nevertheless, the 1992 CSWE Curriculum Policy Statement contains language that is likely to foster a continued emphasis on single-case evaluation, such as in provision M5.7.11, requiring M.S.W. programs to prepare students to "conduct empirical evaluations of their own practice interventions" (CSWE, 1992, p. 5).

So far, the debate about the prospects for greater practitioner use of single-case evaluation methods has tended to focus on motivating more practitioners to use them or finding ways to give practitioners the agency supports and resources they need to use them. This article identifies problems in the analysis of single-case evaluation data that are likely to confront practitioners conducting single-case evaluations. Although some of the technical literature on single-case evaluation discusses data analysis problems (Barlow, Hayes, & Nelson, 1984), the problems to be identified in this article have been neglected as issues influencing the extent to which practitioners are likely to conduct single-case evaluations. The authors postulate that these data analysis problems pose a serious barrier to efforts to promote increased use of single-case evaluation by practitioners and have implications for the way schools of social work are educating practitioners about research.

Some readers might anticipate that this will be another one of the attacks on single-case evaluation and quantitative methods that have appeared in the social work literature since the early 1980s. It is not. The authors value quantitative research and single-case evaluation, and would like to see more of it (just as they would like to see more qualitative research). Our focus is limited exclusively to the prospects for increased use of single-case evaluation *by practitioners*, and whether too much emphasis is being placed on single-case evaluation at the expense of other approaches to *quantitative* evaluation.

AMBIGUOUS DATA PATTERNS

The main data analysis problem to which we refer involves ambiguity in visually interpreting graphed data patterns. Graphs are deemed visually

significant when they clearly imply that the most plausible explanation for improvement in the target problem is the effectiveness of the tested intervention. To rule out the plausibility of alternative explanations (such as history, maturation, cyclicity, and so on), graphs must depict a pattern of unlikely coincidences. These coincidences should show that the target problem begins to improve in a stable manner consistently after the onset of intervention, and not at other times.

Ideally, these unlikely coincidences should be replicated. The more successive replications, the less likely are alternative explanations and the more plausible becomes the hypothesis that the intervention is the real cause of the improvements that occur in the target problem. These replications can be done within one study, such as in ABAB or multiple baseline designs, or across studies. Replicating across studies can entail using a simple AB design for different clients with similar target problems who receive the same intervention.

Although there is a rich literature on successful applications of single-case evaluation by clinical researchers in other fields (Thyer & Thyer, 1992), within-study replications are difficult for social work practitioners to conduct, due to practical constraints, such as heavy caseloads, inadequate agency support, and clients who do not stay in treatment very long (Tolson, 1990). To date, the field has not succeeded in generating widespread social work practitioner use of AB designs without replication, let alone with replication. AB designs are recognized as much more feasible for practitioners to use than are ABAB designs or multiple baseline designs. However, when AB designs are used, we seek to attribute improvement in the target problem to the effects of an intervention based on only one baseline phase and one intervention phase. This is doable, but precarious. To reduce the plausibility of history, cyclicity, or similar alternative explanations in an AB design, the improvement in the B phase must commence almost immediately at the start of that phase and after the end of the A (baseline) phase.

Suppose improvement occurs during the intervention phase, but does not start until long after intervention begins. What would this mean? Would it imply that the intervention was working, but with delayed effects? Perhaps. On the other hand, it might mean that something else happened to cause the improvement. That something might be an extraneous event, perhaps a cyclical change, or something else that coincided with the B phase. Unless the improvement commences almost immediately after the onset of intervention, we cannot call the change a sufficiently dramatic coincidence to enable us to rule out the plausibility of these alternative explanations. There is simply too much time for something else to happen to permit us to conclude that improvement at any point during the B phase means that our intervention

caused the improvement. And yet, it remains entirely plausible that the improvement indeed does reflect the delayed effects of our intervention. The point here is that we simply would not know what to conclude.

If we have conducted a series of replications, then it may not be necessary for improvement in the target problem to commence immediately after the onset of intervention to infer that the intervention is effective. That is, the improvement can commence later during the intervention period. But the timing of the improvement would have to occur in a consistent manner from case to case to establish the principle of unlikely successive coincidences and thereby imply that the most plausible explanation is one of delayed treatment effects (as opposed to the far-fetched speculation that extraneous variables are operating at the same time across different, independent replications).

Single-case evaluation graphs also might seem to have clear implications if virtually no sustained improvement occurs during the intervention phase. This might suggest the ineffectiveness of an intervention. If some improvement appears early during intervention, but then tails off later during intervention, we may wonder whether some extraneous factor is causing just a temporary setback. But even when there is no improvement during the intervention phase, we may wonder whether the study lasted long enough to detect delayed treatment effects.

Thus the types of data patterns that permit practitioners to draw conclusive inferences about their efficacy are rather limited. Based on the authors' experience in conducting single-case evaluations, which we shall illustrate in the next section, we postulate that practitioners will often fail to obtain clear-cut findings as to the effects of their interventions, and that this experience will limit the value they are likely to perceive in continuing to conduct single-case evaluations.

CASE ILLUSTRATION

A study that evaluated the effectiveness of cognitive-behavioral group therapy with male adolescent sex offenders illustrates the types of ambiguous data patterns that can be found in single-case evaluation. That study, which is reported in depth elsewhere (Knox & Rubin, 1994), provided single-case evaluation graphs on 23 offenders, who were distributed among three therapy groups. All the subjects had already been participating in outpatient group therapy that had been court-ordered as part of their probation for a sexual offense in Travis County, Texas. They ranged in age from 13 to 18, with a mean age of 15. Slightly more than half of them were Hispanic; the remainder were about evenly divided between African Americans and Caucasians. None

of them had ever been incarcerated in a correctional facility. Most of the offenses involved siblings or other relatives, and none involved strangers. The age of the victims ranged from 3 to 14 years. Only two of them were repeat offenders.

The baseline on each group started at the same time, and the intervention was introduced in a staggered fashion, producing a multiple baseline design. The baseline lengths for the three groups were 14 days, 21 days, and 28 days. The data were analyzed separately for each of the 23 cases, for the purpose of obtaining numerous replications. Outcome for each case was measured according to whether there was a decrease in antisocial behaviors, an increase in prosocial behaviors or both. Each indicator was self-monitored by the adolescent offenders and reported on a checklist devised for this study. The checklist included 21 antisocial behaviors and 19 prosocial behaviors. The antisocial behaviors ranged from milder forms such as arguing, cussing, and so forth to more extreme forms such as fighting, being sexually inappropriate, being truant, or substance abuse. Where possible, the adolescents' self-reports were triangulated with parental reports of the same behaviors.

For the purpose of this article, we focused on the extent of ambiguity in the study's graphs. We found ambiguous outcomes in 6 of the 23 graphs displaying data on self-monitoring of antisocial behaviors and in 7 of the 23 graphs displaying data on self-monitoring of prosocial behaviors. For the parental reports, we found ambiguous outcomes in 7 of the 16 graphs on antisocial behaviors and 6 of the 16 graphs on prosocial behaviors. The remaining graphs, with only a few exceptions, did not support treatment efficacy. (Because most of our graphs clearly failed to support treatment efficacy, and most of the rest were ambiguous, it can be argued that our study results, overall, were not ambiguous; that is, they unambiguously failed to provide much support for the effectiveness of the intervention. However, it is not our thesis that *our overall* results were ambiguous. Rather, it is our thesis that practitioners conducting single-case evaluation on a single case are likely to come up with an ambiguous graph. As noted above, a different report focuses on the overall report of the study.)

Due to space limitations, we cannot present all of the ambiguous graphs. Instead, we shall present those that best illustrate the data analysis problems to which we refer. Each graph is headed by a "student" number, in which "student" refers to the adolescent in treatment. The graph in Figure 1 (Student 2) shows much less antisocial behavior in the B phase than in the A phase, and the improvement coincides somewhat with the onset of the cognitive-behavioral intervention. This provides some evidence that the intervention may be responsible for the improvement. But this evidence is limited because of the unclear, and somewhat unstable, pattern during baseline—in which the

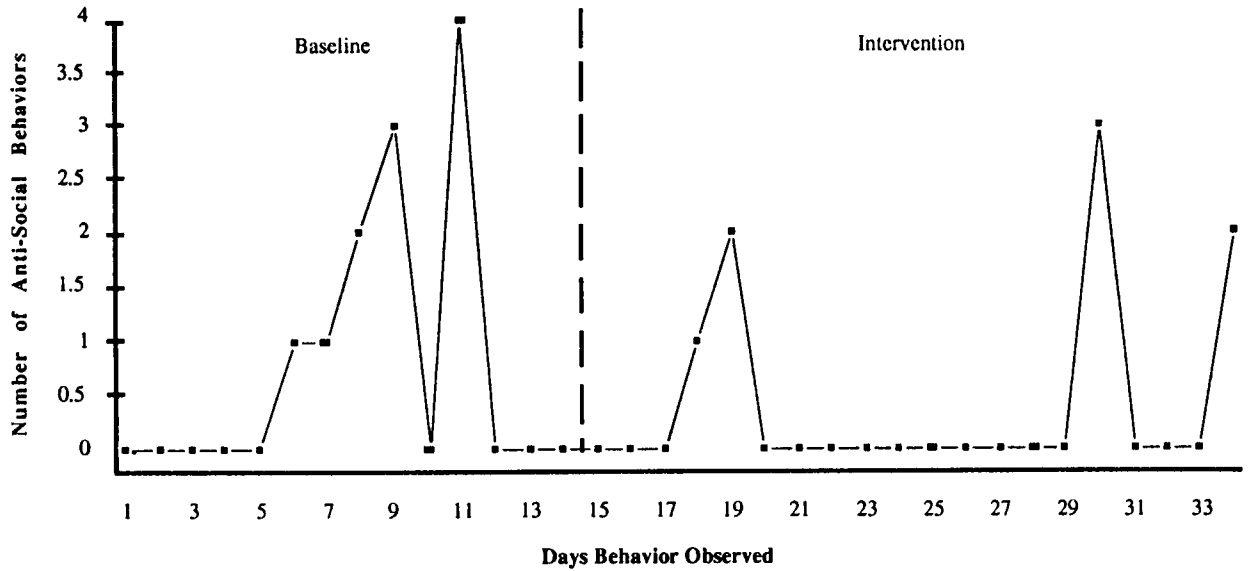

Figure 1: Student 2: Antisocial behaviors.

days when antisocial behavior was reported fall in between periods of no antisocial behavior. Those days, therefore, may represent just a temporary blip up rather than a pattern representative of the adolescent's functioning before intervention.

The graph in Figure 2 (Student 4) also shows much less antisocial behavior in the B phase than in the A phase. But that improvement may have begun on the 10th through 14th days of baseline, and therefore, despite the near absence of antisocial behaviors during B, we cannot conclusively attribute the drop to the effects of the intervention.

The graph in Figure 3 (Student 6) shows a complete disappearance of antisocial behaviors coinciding with the onset of intervention. This, plus the fact that we see no antisocial behaviors during the final 11 days of the study, supports the notion that the intervention may be having powerful effects. But during the middle of the B phase we see a pattern of antisocial behaviors that is not much better (and perhaps is a bit worse) than during baseline. We are therefore left unsure as to whether this is just a cyclical pattern in which the better parts of the cycle just happened to coincide with the beginning and ending weeks of the B phase.

In Figure 4 (Student 12) we see a temporary increase (worsening) of antisocial behaviors at the onset of the intervention, followed by an improvement to levels below baseline at the end of the B phase. What does this mean? One possibility is that the intervention worked. Perhaps the bad start of the B phase reflects just a temporarily adverse reaction to the intervention, or perhaps it reflects a coming out of denial. The stable improvement later on supports the notion that the intervention worked, but that it just took a while for its effects to appear. The alternative, and equally plausible, explanation, however, is that this is an unstable or possibly cyclical pattern. The improvement that started more than midway through the B phase did not coincide with the onset of intervention, and therefore may be due to extraneous factors having nothing to do with the intervention. In other words, we just do not know—the results are ambiguous.

A somewhat similar pattern can be seen for Student 20 in Figure 5. An immediate improvement at the start of B, after a worsening at the end of baseline, may influence some practitioners to suppose that the intervention may be working. But this improvement is followed by a worsening, which in turn is followed by a trend toward improvement at the end of B. Although the improvements at the beginning and end of B offer some hope that the intervention may be effective, the worsening in between suggests that this may simply be a cyclical pattern, having nothing to do with the intervention.

(Text continues on p. 51)

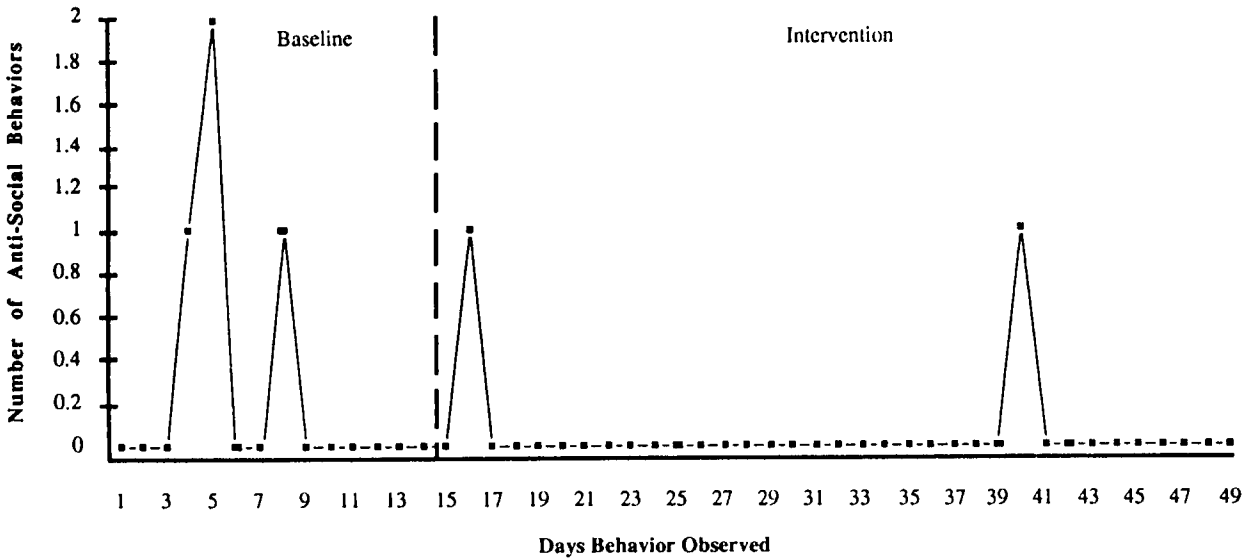

Figure 2: Student 4: Antisocial behaviors.

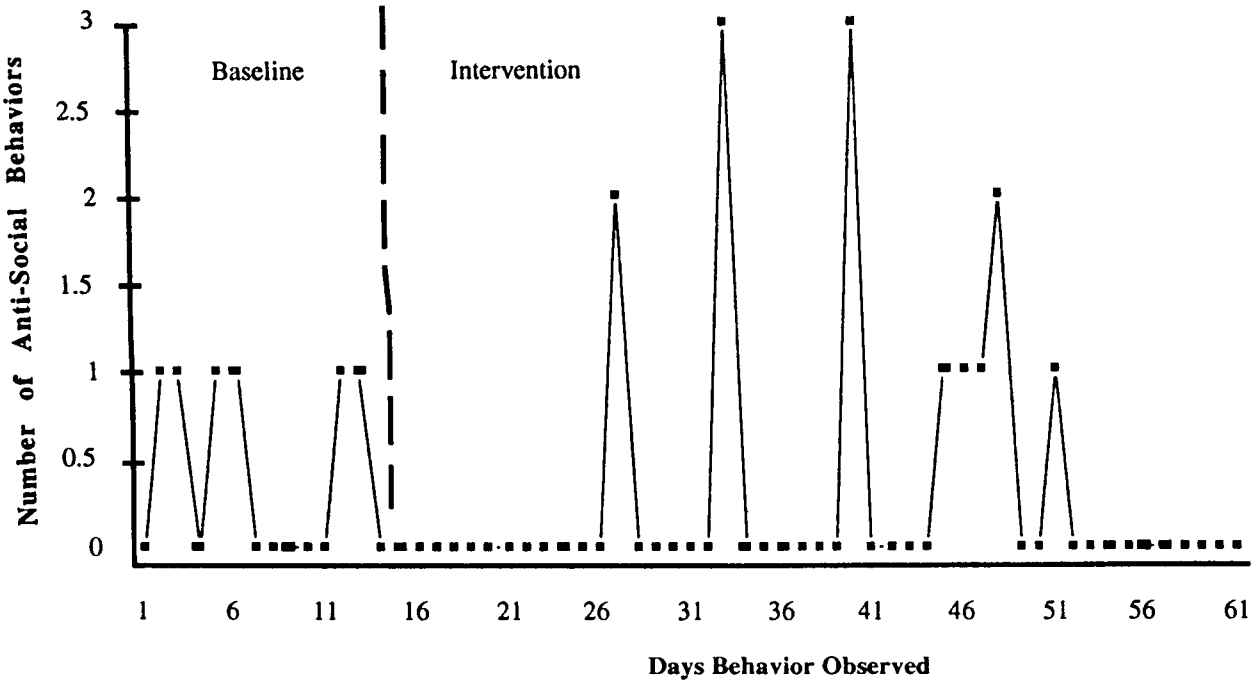

Figure 3: Student 6: Antisocial behaviors.

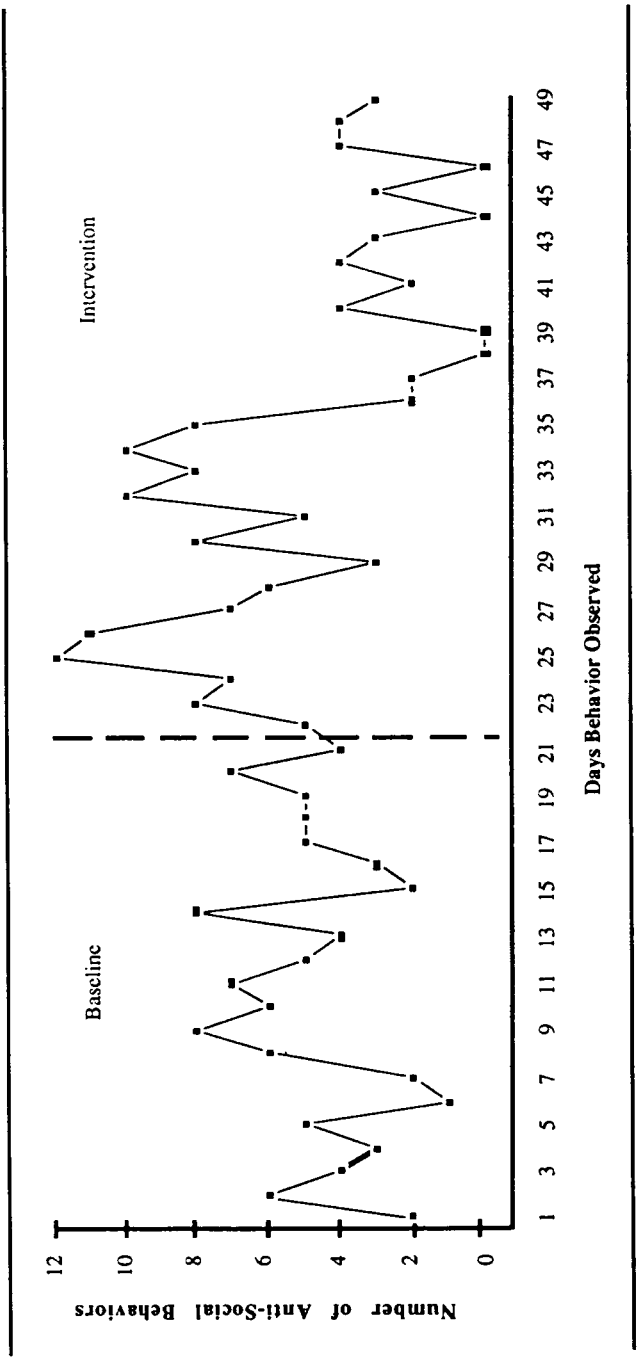

Figure 4: Student 12: Antisocial behaviors.

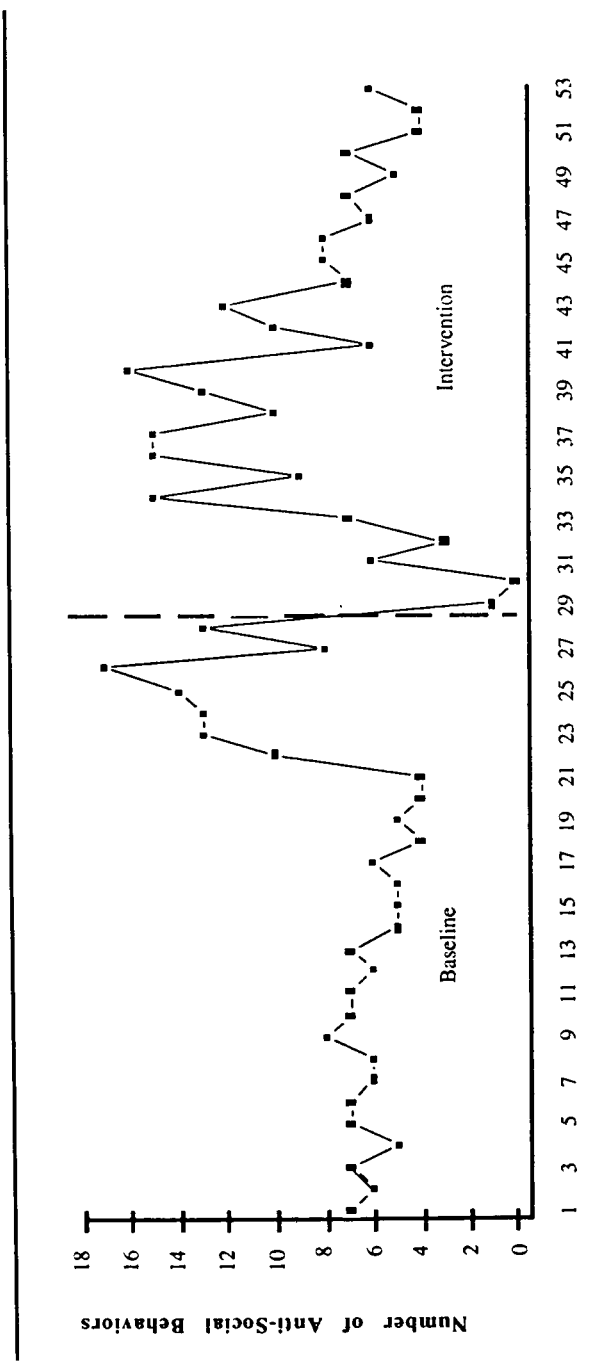

Figure 5: Student 20: Antisocial behaviors.

And yet again, hopeful practitioners might look at the steady improvement at the end of B and wonder whether this would have continued down to zero had the study lasted longer. We do not find this graph to be visually significant. Its pattern does not offer the level of unlikely coincidence needed to conclude that intervention effectiveness is the most plausible explanation. But we think most practitioners obtaining findings like this on their own interventions will find enough hopeful signs in this graph to conclude that its implications regarding their effectiveness are unclear.

Figures 6 and 7 display graphs on prosocial behaviors, concerning whether the intervention was responsible for an increase (i.e., improvement) in prosocial behaviors. The graph in Figure 6 (Student 18) reflects ambiguity regarding a possible cyclical pattern. It is unclear whether the treatment has caused the improvement at the start of the B phase or whether the decline in prosocial behaviors near the middle of the B phase means that a similar decline in the second half of baseline is simply a cyclical fluke, in which the poorer part of the cycle just happened to occur at that time.

The graph in Figure 7 (Student 1) shows an improvement in the frequency of prosocial behaviors commencing at the onset of intervention. But at the end of the intervention phase there is a precipitous decline in the frequency of these desired behaviors. Does this mean that the improvement at the start of intervention can be attributed to a honeymoon period, cyclicity, or some other extraneous factor? Or does it merely reflect a temporary setback during the course of effective treatment, perhaps due to some extraneous event that temporarily disrupted the ongoing trend of improvement? We do not know. Finding the answer would require monitoring the client over a much longer period. Sometimes clinicians (and researchers) can monitor behaviors over much longer periods. Sometimes they cannot.

An additional three graphs showed an increase (i.e., worsening) in self-reported antisocial behaviors throughout the intervention period. Typically, graphs that appear to be visually significant, but in the unanticipated (i.e., harmful) direction, would not be called ambiguous. Instead, they would be interpreted as rather conclusively showing that the intervention was harmful. Yet one of the major short-term treatment objectives of the tested intervention was to overcome the offender's denial and refusal to recognize or admit to his behavior as antisocial. Thus, despite the obvious plausibility of harmful treatment effects (such as through group contagion, for example) as an explanation for these findings, practitioners can also argue that it is plausible that these findings do not mean that the intervention is harmful. That is, they can argue that the results could mean that short-term objectives are being reached and that if the study lasted longer, the antisocial behaviors of these offenders would diminish significantly. It is important to recognize here that

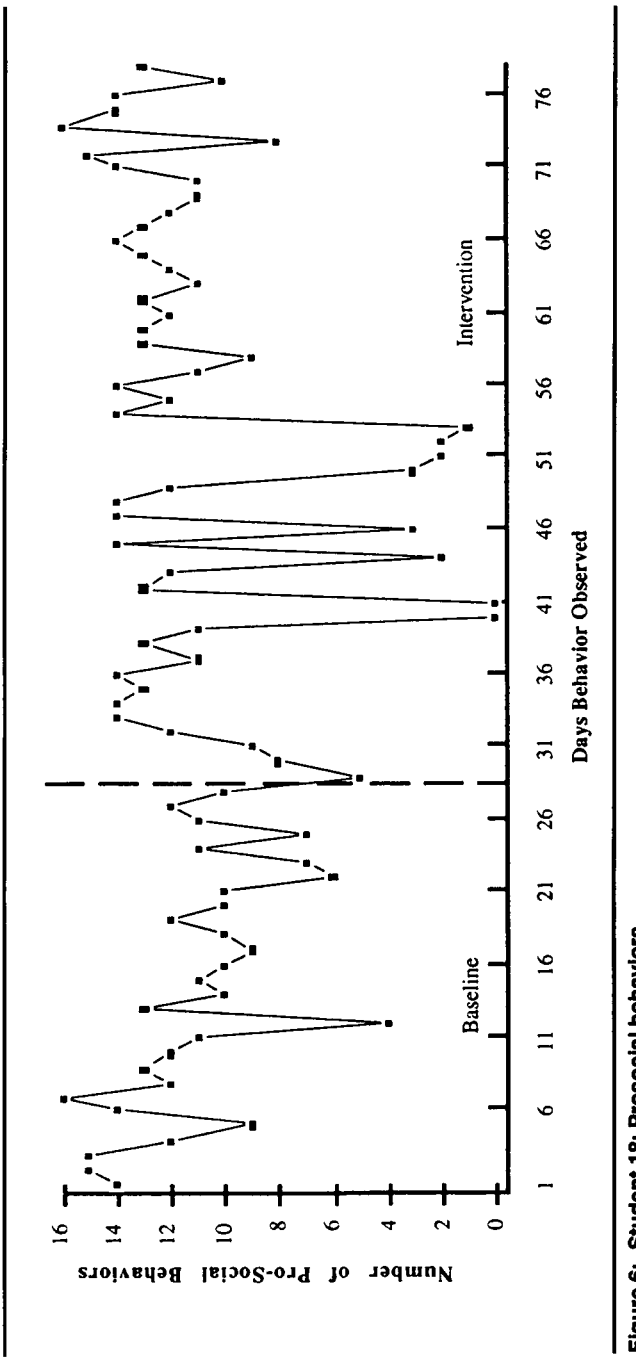

Figure 6: Student 18: Prosocial behaviors.

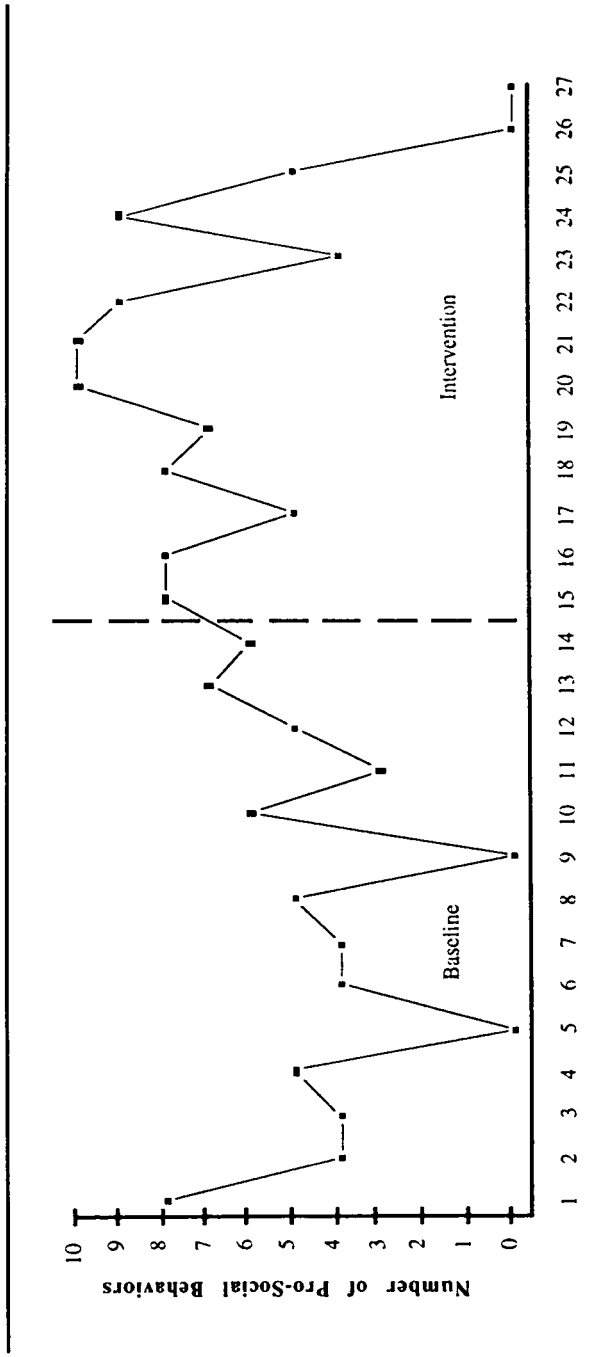

Figure 7: Student 1: Prosocial behaviors.

even if one thinks the practitioners' argument is implausible, it is their *perception* of what these data mean that will influence how valuable they find single-case evaluation. If they perceive data patterns like these to be inconclusive, rather than perceiving them as indicative of a harmful or ineffectual intervention, then regardless of what readers think of their interpretations, those practitioners may become less inclined to employ single-case evaluation. (In the same vein, practitioners might deem as inconclusive those graphs showing no change whatsoever in the intervention period, in that they may wonder whether the study lasted long enough to detect delayed treatment effects.)

STRATEGIES FOR REDUCING AMBIGUITY

One way to attempt to reduce ambiguity and sort out whether the intervention or some alternative explanation seems to be the most plausible explanation is by extending our evaluation of a particular case over a much larger number of data points and then replicating the evaluation over a large number of similar cases. But practitioners are likely to encounter significant practical obstacles to obtaining many data points and many similar cases. Also, as the study discussed above illustrates, replications across many cases may find that improvement commences at very different points after intervention begins for many cases, while many others do not improve.

Consider what would be implied by an evaluation using a one-group pretest-posttest design, in which some cases improved and others did not. We would conclude that the target problems may have improved without treatment, due to a variety of potential factors other than the intervention being evaluated. We would not infer that the intervention was effective just because some of the cases improved. To make such causal inferences, and to rule out the plausibility of alternative explanations, we would need a control group. But in single-case evaluation, we rely on obtaining numerous data points and replication rather than on obtaining a control group. We hope to find unlikely coincidences in which sustained improvement in the target behavior commences at about the same time after the onset of intervention across replications. But what if improvement commences at very different points after the onset of intervention for some cases? And what if many other cases do not improve? Under these conditions we will not have established a sufficiently consistent series of unlikely coincidences to support the argument that the observed improvements are not plausibly due to alternative explanations.

Continuing this line of thought, suppose that a slight majority of cases improve at various points during intervention. It is quite plausible that a

similar percentage of untreated controls may have improved to the same extent. Unlike a single-case evaluation, in a design using a control group we would be able to assess how much improvement occurred among untreated controls and be able to compare it to the extent of improvement among treated cases. If the improvement among treated cases significantly exceeded the improvement among controls, the results would support the efficacy of the evaluated intervention. In other words, the *precise timing* of the onset of improvement from case to case would not matter. Neither would it matter whether all or even most of the treated cases improved. Even if most of the treated cases did not improve, treatment efficacy might be indicated if a much smaller proportion of untreated controls improved. But in single-case evaluation we lack comparison rates. Therefore, we are forced to look for dramatic unlikely coincidences in the *timing* of *stable* improvement. Moreover, these coincidences must replicate consistently from case to case.

How reasonable is the implicit expectation in single-case evaluation that when evaluating effective interventions dramatic coincidences in the timing of stable improvement can be found across cases? With some types of target populations and target problems that expectation may seem realistic, such as when behavioral modification and social skills training interventions are applied to problems that respond quickly to them, and, as noted earlier, there is a rich literature in allied fields showing this to be so. But social workers often work with problems or populations that do not respond so quickly to intervention. For example, in a study of an outreach group work program for battered women, Rubin (1991) found ambiguous data patterns similar to the ambiguity discussed in the graphs illustrated previously, patterns in which questions about delayed treatment effects and overcoming denial complicated the interpretation of graphs across cases. Numerous other types of interventions and target populations would seem to pose similar concerns about the likelihood of obtaining unambiguous graphed data patterns. For example, consider play therapy interventions with children who have been abused or the treatment of adult survivors of sexual abuse. In light of variability across cases in overcoming denial, in building trust, and so on, it does not seem realistic, even when we are assessing effective interventions, to suppose that single-case evaluations will consistently yield graphs showing stable improvement coinciding with the onset of treatment or commencing at about the same time from case to case. Neither does it seem particularly realistic to suppose that practitioners will be able to sustain the duration of monitoring nor the extent of replications needed to sort out intervention effects when results are ambiguous and inconsistent.

RELYING ON STATISTICAL SIGNIFICANCE TESTING

Another strategy that has been proposed to offset visual ambiguity is through the use of statistical techniques designed for practitioners conducting single-case evaluations. Those advocating this strategy argue that testing whether the data in the graph are statistically significant can be used as the criterion for deciding whether visually ambiguous results support treatment efficacy. But this argument has several problems. One is that statistical significance refers only to ruling out chance, not to determining whether it was the intervention or some extraneous factor that caused the statistically significant change. For example, in an uncontrolled pre-experimental design we would not conclude that the intervention caused an improvement in the target problem just because there was a statistically significant improvement from pretest to posttest. Despite our ability to rule out chance as the explanation for the improvement, we would recognize the need for a more internally valid design to draw causal inferences about intervention effects, and we would not let statistical conclusion validity obviate our doubts about internal validity. For the same reason, a statistically significant improvement during the B phase of an AB design does not obviate the need for unlikely coincidences in the timing and stability of the improvement.

Another problem is the low statistical power inherent in single-case evaluations. Texts on this topic designed for social workers suggest that practitioners often find it unfeasible to obtain more than 10 data points in each phase of an AB design (Bloom & Fischer, 1982; Rubin & Babbie, 1993). The power of a two-tailed significance test with a significance level of .05, assuming a moderately effective intervention, is only about .25 in an AB design with 10 data points in each phase (Rubin & Babbie, 1993). This means that we take a huge risk of a Type II error when we conduct significance testing in such studies. If practitioners understand this, then they will know that statistical significance tests are unlikely to help them develop more conclusive, clinically useful implications from visually ambiguous data patterns.

Moreover, the statistical techniques that have been designed for practitioners conducting single-case evaluations can come up with results that are contradictory and which defy logic. Consider, for example, the data in the graph in Figure 8, which come from the sex offender treatment evaluation study discussed previously. We see a visually ambiguous data pattern regarding the intervention's effect on antisocial behaviors, as discussed earlier regarding Figure 5. Using the two standard deviation procedure to test for statistical significance in this graph, we would find a baseline mean of 7.21.

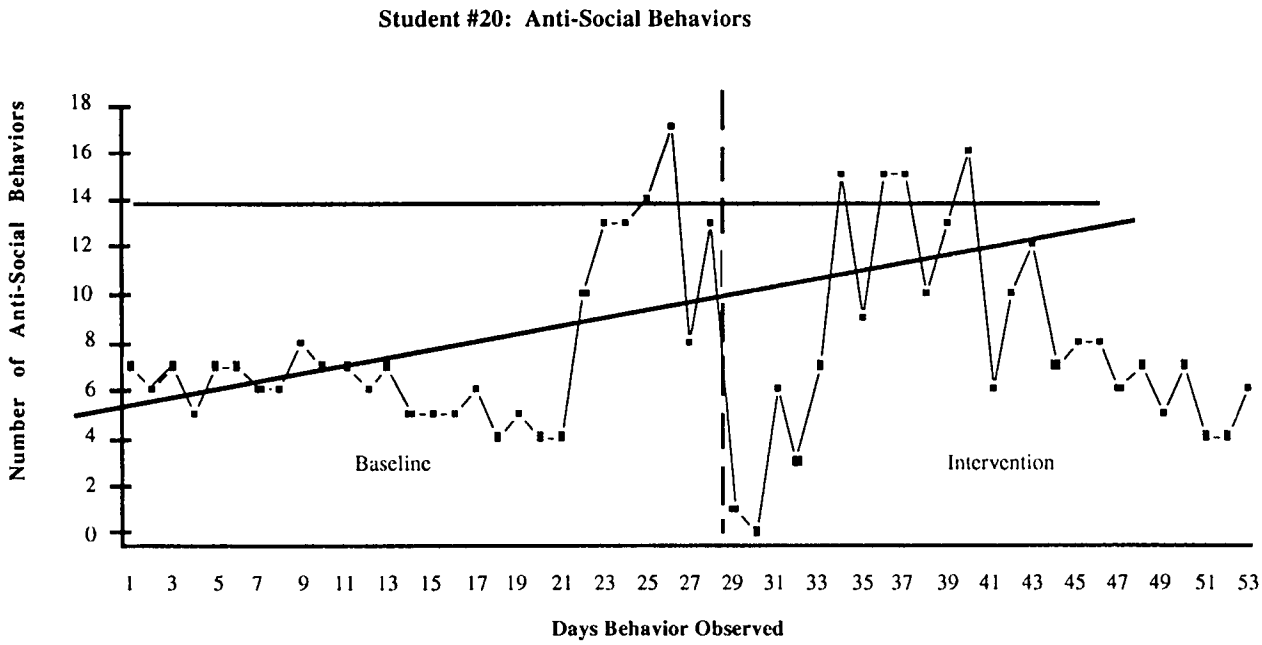

Figure 8: Graph displaying a visually ambiguous data pattern, statistically significant results in the undesirable direction using the two standard deviation procedure, and statistically significant results in the opposite (desired) direction using the celeration line approach.

Two standard deviations equals 6.65. Two consecutive data points during intervention (points 36 and 37) fall more than two standard deviations above the baseline mean (i.e., they are above 13.86). Thus, they are in the undesirable zone, because the desired effect is a reduction in antisocial behaviors. Because Bloom and Fischer (1982) recommend as a criterion for significance finding two consecutive data points during intervention that fall more than two standard deviations away from the baseline mean, relying on this significance test to resolve the graph's visual ambiguity would influence us to conclude that the intervention had statistically significant harmful effects. Yet the visual pattern in this graph shows that such a conclusion would be ludicrous, because a similarly bad cycle can be found at the end of baseline, and because the remaining data during intervention show a consistently improving trend.

Furthermore, using the celeration line approach to test for the statistical significance of these data would yield the opposite conclusion. That is, the proportion of data points in the desired zone (under the celeration line) during intervention is significantly greater than the proportion during baseline, indicating that the intervention had statistically significant beneficial effects. This, too, would not be logical. A visual analysis reveals that the only reason that the celeration line approach yields statistical significance is the blip up at the end of baseline. Because of this blip, practitioners might be unsure as to which of the foregoing two statistical significance approaches to use with these data. Imagine their reaction if they tried both approaches and found that they yield contradictory conclusions.

Of course, inconclusive statistical results can be found in all kinds of research, not just single-case evaluation. But the focus of this article is on *practitioners* producing single-case evaluations. In that proponents of single-case evaluation, such as Bloom and Fischer (1982), often find it unfeasible to obtain more than 10 data points in each phase of an AB design, it seems reasonable to suppose that practitioners are particularly vulnerable to having low statistical power, and therefore inconclusive results, when they conduct single-case evaluation. Moreover, the problems (illustrated previously) in the statistical procedures developed for practitioners doing single-case evaluation further suggest that this kind of evaluation is more likely to produce inconclusive findings than are other types of evaluation.

COUNTERARGUMENTS

The foregoing argument is built around the need to obtain clear-cut results that permit causal inferences. Some may counter that single-case evaluations

with ambiguous results can have value as exploratory steps in the process of accumulating many replications and using inductive logic to tease out working hypotheses about the conditions required for a particular intervention to be effective and how long it takes for its effects to begin appearing. We would agree, but would argue that it is more realistic to expect practice-based researchers to carry out that very long-range and time-consuming research agenda than it is to expect practitioners to do it. The extensive resources and data-gathering time that would be required would probably far exceed that required of many research strategies involving traditional group designs. Is it realistic to suppose that practitioners will find its elusive long-range payoff sufficient inducement to motivate them to use single-case techniques to evaluate their own practice?

Another counterargument is that practitioners may find value in collecting and graphing repeated measurements because it gives them more clinical information and a better empirical basis for assessing when a target problem has been adequately alleviated or when a different intervention should be tried. But visually ambiguous graphs complicate such clinical decisions. If it is unclear from the graph whether the intervention is working, or has had enough time to work, it is difficult to rely on the data for clinical decisions about when to abandon the intervention and try something new. Likewise, without clear, stable data patterns collected over a long period, it is unclear whether the graphed improvement in the target problem is only a temporary cycle.

A related counterargument maintains that single-case designs serve a sufficient purpose if they merely show, from a clinical standpoint, that adequate change has taken place, regardless of what caused it. Thus it can be argued that the fact that graphs can show whether target behaviors seem to have been attained is a sufficient benefit to augur well for the prospects of increased practitioner conducting of single-case evaluation. We would note, however, that this argument represents a vision of the benefits of single-case evaluation that is considerably scaled down from the original vision of single-case evaluation, the vision that captured the imagination of many social work educators by promoting single-case evaluation as a way to evaluate one's practice effectiveness. Moreover, if single-case evaluation means little more than monitoring whether there is any change in clients, and does not focus on attempting to ascertain what caused the change, then this raises questions about whether it really represents a significant way to integrate practice and research. This vision of single-case evaluation seems to see it more as a clinical assessment tool than as a way for practitioners to answer research questions about their own practice.

Perhaps the most seductive counterargument (Downs, 1994) to this article's thesis is to assert that the interventions social workers are using have already been shown to be effective in group level evaluations and to argue that therefore practitioners do not need unambiguous, conclusive, visually significant graphs to rule out the plausibility of alternative explanations for the data patterns they observe. If the interventions they are using have already been shown to be effective in group studies, practitioners would not need to worry as much about the timing of the improvement. The improvement could commence at almost any point during intervention. As long as the data pattern was merely consistent with the plausibility of intervention efficacy, practitioners might conclude that the intervention appeared to be working with this particular case. This argument presumes that practitioners would not be conducting single-case evaluations to test the general effectiveness of an intervention; instead, they would use them to see whether the probabilistic findings of group research supporting the effectiveness of interventions apply to a particular case.

This counterargument correctly notes that group research gives us only probabilities, that each case-practitioner dyad is unique, and that different cases will respond to tested interventions in different ways. But if this is so, then it would seem to support this article's point that single-case evaluations of effective interventions are not likely to produce graphs that are consistent across cases in the timing of improvement. Therefore, the strength of this counterargument would appear to depend on whether the interventions that social workers use have already had sufficient group research support of their efficacy. Some interventions, particularly some behavioral ones, have had considerable support. But many others have not. Moreover, even if practitioners assume they are applying an intervention that group research has found to be effective, they may have to further assume that the beneficial effects will emerge on their graphs near the onset of intervention. If the treatment effects are delayed, then they may need to conduct lengthy monitoring to ensure that they do not prematurely decide that the intervention is not working with their particular case. How realistic is it to suppose that practitioners will want to continue this type of data collection from case to case? If they are assuming that group research has already established the efficacy of their intervention, are they not more likely to rely on traditional forms of clinical feedback to make decisions about how well particular clients are responding to the intervention?

IMPLICATIONS FOR SOCIAL WORK PRACTICE AND EDUCATION

The foregoing analysis suggests the need to reconsider whether substantial increases in the conducting of single-case evaluation among social work practitioners can be achieved and sustained through educational or motivational strategies. It suggests that although the possibility of obtaining inconclusive findings inheres in all research and evaluation designs, the likelihood of obtaining inconclusive findings is particularly problematic for social work practitioners using single-case designs to evaluate their own, nonbehavioral forms of practice. We postulate two reasons why it is particularly problematic for them: (a) because they are more likely to obtain inconclusive findings than are researchers employing single-case or other designs; and (b) because inconclusive findings are more likely to sap practitioner enthusiasm for sustained investment in employing single-case evaluation than they are to sap researcher enthusiasm for sustained involvement in conducting research. If this thesis is correct, then efforts to increase agency supports and resources for single-case evaluation may, in the long run, yield disappointing returns, as those practitioners and administrators who are persuaded to invest in single-case evaluation encounter findings that do not give them the degree of guidance for practice that they expected.

Likewise, our thesis suggests that the current emphasis on single-case evaluation in the social work curriculum may be overdone. This in turn implies the need to reconsider the wording of CSWE curriculum policy and accreditation standards that may be influencing schools of social work to pursue a heavy emphasis on single-case evaluation in both their research and practice curricula. Reducing the amount of emphasis on single-case evaluation in the research curriculum would free up time to try to enhance students' grasp of other research content areas. Some M.S.W. programs also may want to reduce the emphasis on single-case evaluation in the practice curriculum. For example, like some other M.S.W. programs, the master's program with which the authors are associated requires a large portion of the second-year field practicum and practice seminar to be devoted to conducting a single-case evaluation on a case the student has in the field. Deleting that requirement would provide more room to work on improving practice competence, which some research suggests is sorely in need of improvement (Rubin, Franklin & Selber, 1992), or on using the research of others to guide practice.

We are not arguing for less research curriculum or for less research-practice curriculum integration; we are just questioning whether those curricular areas are overrelying on single-case evaluation to achieve their aims. Neither are we recommending that single-case evaluation be deleted from social work education. Some researchers have published useful single-case evaluation studies (Thyer & Thyer, 1992), most commonly but not exclusively involving behavioral and/or cognitive interventions. (We wonder how many other single-case evaluation studies produced findings that were so ambiguous that their investigators deemed them not worth submitting for publication.) We believe that students should learn about using or producing single-case evaluation studies to the same extent to which they learn about useful surveys, group experiments, quasi-experiments, and so on. In other words, rather than going overboard in emphasizing single-case evaluation as some sort of panacea for integrating research and practice, we should treat it the same way we treat other research methods and designs.

A second, and different, implication of this analysis pertains to *how* we teach students or social work practitioners about single-case evaluation. We should make sure they are taught about the foregoing data analysis problems. We should not oversell single-case evaluation or, by neglecting data analysis problems, implicitly foster the notion that single-case evaluations are likely to produce conclusive findings that will provide clear implications for practice. We should prepare students and practitioners to encounter and deal with these data analysis problems. Forewarning them and forearming them about these problems seems more likely to help them sustain a commitment to conducting single-case evaluations in the long run than is neglecting these problems.

Likewise, we should not teach in a simplistic, mechanistic fashion the statistical significance tests designed for practitioners conducting single-case evaluations. Neither should we imply that the results of these tests can be relied on to determine which rival explanations are to be inferred from visually ambiguous graphs. By using data like those displayed in Figure 8, we should demonstrate the hazards of such reliance. Practitioners and students should understand the large risk of Type II errors taken in many single-case evaluation studies, and they should understand the difference between internal validity and statistical conclusion validity. Regardless of whether we are teaching about group research or single-case evaluation, we should make sure students understand that if design considerations do not permit drawing conclusive causal inferences, statistical significance should not overrule design considerations in interpreting causality, regardless of how statistically significant a study's findings are.

Practitioners and students should also learn that the data analysis problems inherent in single-case evaluations can require long-range, labor intensive, inductive research efforts to sort out when and under what conditions certain interventions are effective. This applies particularly to target problems that are not likely to respond immediately to effective interventions. Finally, this analysis implies the importance of teaching students about group research studies and how to critique them. This implication is particularly important if a heavy emphasis on single-case evaluation is retained, based on the assumption that group research has already demonstrated the probabilistic efficacy of social work interventions. Students should be prepared to understand and to critically evaluate that assumption.

Social work has had a long history of going overboard in embracing and overemphasizing new research modalities as panaceas, dating back to the social survey movement at the turn of this century (Zimbalist, 1977). Eventually, as the limitations of these modalities become better understood, we teach about them in a more realistic, critical manner, giving them about the same emphasis as other research modalities that have both advantages and disadvantages. The single-case evaluation movement seems to fit this historical pattern.

REFERENCES

Barlow, D., Hayes, S., & Nelson, R. (1984). *The scientist practitioner: Research and accountability in clinical and educational settings*. New York: Pergamon.

Bloom, M., & Fischer, J. (1982). *Evaluating practice: Guidelines for the accountable professional*. Englewood Cliffs, NJ: Prentice-Hall.

Blythe, B. (1983). *An examination of practice evaluation among social workers*. Unpublished doctoral dissertation, University of Washington, Seattle.

Blythe, B. (1990). Improving the fit between single-subject designs and practice. In L. Videka-Sherman & W. J. Reid. (Eds.), *Advances in clinical social work research* (pp. 29-32). Silver Spring, MD: National Association of Social Workers.

Briar, S. (1990). Empiricism in clinical practice: Present and future. In L. Videka-Sherman & W. J. Reid (Eds.), *Advances in clinical social work research* (pp. 1-7). Silver Spring, MD: National Association of Social Workers.

Corcoran, K. J. (1990). Illustrating the value of practice wisdom. In L. Videka-Sherman & W. J. Reid (Eds.), *Advances in clinical social work research* (pp. 54-57). Silver Spring, MD: National Association of Social Workers.

Council on Social Work Education. (1992). *Curriculum policy statement for master's degree programs in social work education*. Alexandria, VA: Author.

Dean, R., & Reinherz, H. (1986). Psychodynamic practice and single system design: The odd couple. *Journal of Social Work Education, 22*, 71-81.

Dolan, M. M., & Vourlekis, M. M. (1983). A field project: Single-subject design in a public social service agency. *Journal of Social Service Research, 6*, 29-43.

Downs, W. R. (1994). Lacking evidence of effectiveness, should single-case evaluation techniques be encouraged in practice? Yes. In W. W. Hudson & P. S. Nurius (Eds.), *Controversial issues in social work research* (pp. 113-119). Boston, MA: Allyn & Bacon.

Fortune, A. E. (1982). Teaching students to integrate research concepts and field performance standards. *Journal of Education for Social Work, 18*(1), 5-13.

Gingerich, W. J. (1977). The evaluation of clinical practice: A graduate level course. *Journal of Social Welfare, 4*, 109-118.

Gingerich, W. J. (1984). Generalizing single-case evaluation from classroom to practice setting. *Journal of Education for Social Work, 20*(1), 74-82.

Gingerich, W. J. (1990). Rethinking single-case evaluation. In L. Videka-Sherman & W. J. Reid (Eds.), *Advances in clinical social work research* (pp. 11-24). Silver Spring, MD: National Association of Social Workers.

Ivanoff, A., Blythe, B., & Briar, S. (1987). The empirical clinical practice debate. *Social Casework, 65*, 290-298.

Jayaratne, S., & Levy, R. (1979). *Empirical clinical practice*. New York: Columbia University Press.

Knox, K., & Rubin, A. (1994, March). *Cognitive-behavioral group therapy with adolescent sex offenders*. Paper presented at the annual meeting of the Council on Social Work Education, Atlanta, GA.

Mutschler, E. (1984). Evaluating practice: A study of research utilization by practitioners. *Social Work, 29*, 332-337.

Nelsen, J. C. (1981). Issues in single-subject research for nonbehaviorists. *Social Work Research & Abstracts, 17*, 31-37.

Nelsen, J. C. (1990). Single-case research and traditional practice: Issues and possibilities. In L. Videka-Sherman & W. J. Reid (Eds.), *Advances in Clinical Social Work Research* (pp. 37-47). Silver Spring, MD: National Association of Social Workers.

Penka, C., & Kirk, S. (1991). Practitioner's involvement in clinical evaluation. *Social Work, 36*, 513-518.

Reinherz, H. (1990). Beyond regret: Single-case evaluations and their place in social work education and practice. In L. Videka-Sherman & W. J. Reid (Eds.), *Advances in clinical social work research* (pp. 25-28). Silver Spring, MD: National Association of Social Workers.

Richey, C. A., Blythe, B. J., & Berlin, S. B. (1987). Do social workers evaluate their practice? *Social Work Research & Abstracts, 23*, 14-20.

Robinson, E.A.R., Bronson, D., & Blythe, B. J. (1988). An analysis of the implementation of single-case evaluation by practitioners. *Social Service Review, 62*, 285-301.

Rubin, A. (1991). The effectiveness of outreach counseling and support groups for battered women: A preliminary evaluation. *Research on Social Work Practice, 1*, 332-357.

Rubin, A., & Babbie, E. (1993). *Research methods for social work* (2nd ed.). Pacific Grove, CA: Brooks/Cole.

Rubin, A., Franklin, C., & Selber, K. (1992). Integrating research and practice into an interviewing skills project: An evaluation. *Journal of Social Work Education, 28*, 141-152.

Siegel, D. H. (1983). Can research and practice be integrated in social work education? *Journal of Education for Social Work, 19*, 12-19.

Simons, R. (1987). The impact of training for empirically based practice. *Journal of Social Work Education, 23*, 24-30.

Stern, S. B. (1990). Single-system designs in family-centered social work practice. In L. Videka-Sherman & W. J. Reid (Eds.), *Advances in clinical social work research* (pp. 48-53). Silver Spring, MD: National Association of Social Workers.

Thyer, B. A. (1990). Single-system research designs in social work practice. In L. Videka-Sherman & W. J. Reid (Eds.), *Advances in clinical social work research* (pp. 33-36). Silver Spring, MD: National Association of Social Workers.

Thyer, B. A., & Thyer, K. B. (1992). Single-system research designs in social work practice: A bibliography from 1965 to 1990. *Research on Social Work Practice, 2*, 99-116.

Tolson, E. R. (1990). Why don't practitioners use single-subject designs? In L. Videka-Sherman & W. J. Reid (Eds.), *Advances in clinical social work research* (pp. 58-64). Silver Spring, MD: National Association of Social Workers.

Welch, G. J. (1983). Will graduates use single-subject designs to evaluate their casework practice? *Journal of Education for Social Work, 19*, 42-47.

Zimbalist, S. (1977). *Historic themes and landmarks in social welfare research.* New York: Harper & Row.