The Journal of Experimental Education, *2006*, *75*(1), 66–81

Tests for the Visual Analysis of Response-Guided Multiple-Baseline Data

JOHN FERRON
PEGGY K. JONES
University of South Florida

ABSTRACT. The authors present a method that ensures control over the Type I error rate for those who visually analyze the data from response-guided multiple-baseline designs. The method can be seen as a modification of visual analysis methods to incorporate a mechanism to control Type I errors or as a modification of randomization test methods to allow response-guided experimentation and visual analysis. The approach uses random assignment of participants to intervention times and a data analyst who is blind to which participants enter treatment at which points in time. The authors provide an example to illustrate the method and discuss the conditions necessary to ensure Type I error control.

Key words: multiple-baseline, randomization test, response-guided, single-case, visual analysis

VARIABILITY IN BASELINE DATA leads to difficulties in making treatment effect inferences. To avoid these difficulties, researchers may try to identify and control the sources of variation. If completely successful, the researcher will have demonstrated control over the behavior, obtained a constant baseline, and will be in a position where it is relatively easy to assess treatment effects. Applied researchers, however, will often find that there are factors outside their control and that a constant baseline cannot be obtained. These researchers will often turn to other strategies to facilitate inference, such as increasing the number of observations during baseline or adding design elements (e.g., additional baselines or reversals). These strategies tend to be helpful, but when baseline data are variable, different analysis strategies may still lead to different conclusions.

Address correspondence to: John Ferron, Educational Measurement and Research, University of South Florida, 4202 East Fowler Ave., EDU 162, Tampa, FL 33620. E-mail: ferron@tempest.coedu.usf.edu

In choosing among data analysis strategies, researchers should be clear about the inferences they wish to make. Different inferences will call for different analyses. Robinson and Levin (1997) provided a framework for making effect inferences where researchers first consider whether the treatment had an effect. If an effect can be inferred, researchers then move to more complex inferences about the size of the effect and the practical (or clinical) significance of the effect. In this article, we focus on the inference of whether the treatment had an effect because we see it as a critical first step. If one cannot provide a convincing argument that the treatment had an effect, it does not seem possible to make a convincing argument that the treatment effect was sizable or that it was clinically significant.

When making inferences that a treatment had an effect, there are two possible types of errors: (a) concluding a treatment had an effect when it did not (Type I) and (b) failing to conclude a treatment had an effect when it did (Type II). When selecting among methods for making inferences that a treatment had an effect, one typically identifies methods that control the Type I error rate. From among the methods that control the Type I error rate, one then selects the method that is most sensitive to finding effects (the one with the lowest Type II error rate).

How well a particular analysis for single-case data controls the Type I error rate depends on whether or not the design is response-guided (Allison, Franklin, & Heshka, 1992; Ferron, Foster-Johnson, & Kromrey, 2003; Todman & Dugard, 1999). In a response-guided design, the number of observations made in each phase depends on the emerging data. For example, a researcher who intervenes only after baseline data have stabilized is using a response-guided design. A researcher who waits until a participant in treatment reaches criterion before intervening with another participant is also using a response-guided design. When designs are not response-guided, the number of observations in each baseline and series is determined prior to collecting data. Although we are primarily concerned with response-guided designs, we provide background by first considering Type I error rates when designs are not response-guided.

Type I Error Rates When Designs Are Not Response-Guided

Visual Analyses

Several researchers (Fisch, 2001; Fisher, Kelley, & Lomas, 2003; Matyas & Greenwood, 1990; Stocks & Williams, 1995) have examined Type I error rates for visual analyses when designs are not response-guided. Each of these researchers found that there are conditions under which visual analysts are too likely to see effects in graphs of data that are generated without treatment effects. Averaging across the zero effect size conditions in the Matyas and Greenwood study led to an average Type I error rate estimate of .24, with the highest estimates resulting from conditions based on data with an autocorrelated error structure. The average estimated Type I error rate in the Stocks and Williams study was .25, with the higher estimates coming from conditions where the data generation included a trend. Fisch summarized four studies, all of which showed elevated Type I error rates. Averaging across conditions within studies and then across studies yielded an average estimate of .28. Finally, Fisher, Kelley, and Lomas estimated an average Type I error rate of .24 prior to training participants to use a more structured method of conducting visual analysis.

Structured Visual Analyses

Training methods have been suggested as a method for potentially improving the accuracy of visual analysts (Parsonson & Baer, 1992). By using training and structured criteria, researchers have been able to demonstrate that agreement among visual analysts can be improved (Hagopian et al., 1997) and that the accuracy of visual analysts can be improved (Fisher et al., 2003). A dual criteria (DC) method was found to have Type I error rates that exceeded .05 for some of the conditions studied, which motivated the development of a conservative dual criteria (CDC) that produced Type I error rates under .05 for each of the conditions studied (Fisher et al.). It should be noted that the Type I error rates for the CDC did vary across conditions, which were obtained by varying the series lengths (10 and 20) and varying the level of autocorrelation (0 to .5). Information is not available on the Type I error rate of these methods when the data contain trends (linear or nonlinear) or when the error structure is more complex (e.g., second-order autoregressive).

Statistical Modeling

Analyses based on statistical models for time series data have also been shown to control the Type I error rate under certain assumptions. For example, regression analyses can be considered if one assumes that the errors are independent (Huitema & McKean, 1998). There has been considerable debate about the plausibility of this assumption and the feasibility of using the data from single-case studies to test for independence (e.g., Huitema, 1985; Huitema & McKean, 2000; Kratochwill et al., 1974; Matyas & Greenwood, 1997). If one cannot assume independence but can assume a first-order autoregressive model, then a double-bootstrap method is available that has been shown to do a relatively good job of controlling the Type I error rate (McKnight, McKean, & Huitema, 2000).

For researchers who believe that the errors may be dependent in a manner that cannot be accurately summarized by a first-order autoregressive model, more general time series analysis models (Box, Jenkins, & Reinsel, 1994) may be considered. These more general models require that analysts select among many possible error structures, which requires identification procedures that tend to be unreliable with series lengths of 50 and less (Velicer & Harrop, 1983; Westra, 1979). Furthermore, incorrect identification leads to distorted Type I error rates

(Greenwood & Matyas, 1990; Padia, 1976). Consequently, these models are generally not recommended unless there is a minimum of 50 to 100 observations in the series (Box et al.; McKnight et al., 2000), which reduces their utility for many areas of behavioral research.

Randomization Tests

Randomization tests have been shown to control the Type I error rate when the design incorporates randomization, the randomization distribution is based on the randomization strategy used, and the test statistic is chosen independently of the data (Edgington, 1980). Under these conditions, the Type I error rate is controlled regardless of the complexity of the data. When designs do not contain randomization, the ability of randomization tests to control Type I error rates can no longer be demonstrated mathematically. Results of a Monte Carlo study of Type I error rates under conditions of nonrandom assignment suggests that there are many conditions for which randomization tests fail to control the Type I error rate (Ferron et al., 2003).

Response-Guided Designs

Response-guided experimentation gives researchers the flexibility to extend phases based on the pattern seen in the observed data (e.g., gathering additional baseline data to obtain stability). This flexibility allows researchers to capitalize on chance. More specifically, it allows for repeated visual tests of the data as they accumulate (Allison et al., 1992), and it is known that the repeated testing of accumulating data increases the Type I error rate (Armitage, McPherson, & Rowe, 1969). In a study examining visual analyses, Todman and Dugard (1999) found that analysts were more likely to see an effect in chance data when interventions were designated after a relatively stable sequence of observations. In a study of the impact of response-guided experimentation on randomization tests, Ferron et al. (2003) demonstrated that Type I error rates were inflated under many conditions and that the error rates depended on the complexity of the data and the algorithm used to simulate the response-guided process. Although we are not aware of studies that have directly assessed the effect of response-guided experimentation on structured visual analyses or statistical modeling, one would suspect that these analyses would also have difficulties because they were not designed with mechanisms to account for the potential biases created by response-guided experimentation. In sum, there are no methods for analyzing data from response-guided multiple-baseline designs for which Type I error control has been demonstrated.

One strategy for dealing with this problem would be to avoid response-guided experimentation. To get a better sense of the degree to which multiple-baseline researchers rely on response-guided experimentation, we conducted a survey of multiple-baseline applications using the key word multiple baseline in the PsycINFO and ERIC databases. Of 101 concurrent multiple-baseline studies identified between 1998 and 2001, we found that 31 (31%) of the studies contained an explicit statement of response-guided experimentation. Common examples included stating that the treatments started after the baseline data had stabilized and that treatment phase data were allowed to reach some criterion before intervening with the next participant. Another 52 (51%) appeared to be response-guided. The authors of these studies did not explicitly state how the interventions were assigned, but the spacing of the interventions was irregular, which tends to occur in response-guided applications. There were 13 (13%) that contained an explicit statement that the interventions were established in some systematic way prior to observing the data, and the remaining 4 (4%) appeared systematic in that the spacing between interventions was equal. In none of the studies reviewed was there any evidence that the researcher established the placement of the intervention points randomly. Agreement among raters classifying these articles was 95%, computed as the number of agreements divided by the total number of decisions.

One explanation for these findings is that researchers are not aware of the analysis difficulties posed by response-guided experimentation and they are not aware that these difficulties can be resolved through randomization. Another explanation is that response-guided experimentation offers advantages that are judged to outweigh the difficulties. Response-guided experimentation allows researchers to extend baseline phases when there is more variability than anticipated, a trend, or a troubling outlier, and to extend treatment phases if the behavior is more variable than expected during treatment, or if the effect is delayed, occurs gradually, or is relatively small. These extensions lead to more data and presumably greater chances of seeing effects (fewer Type II errors). In addition, the extensions help researchers resolve uncertainties about trends, which may facilitate more complex inferences regarding effect size or practical significance. For these inferences, one must rely heavily on baseline projections of what would have happened had there been no treatment.

The perceived value of response-guided experimentation coupled with the desire to control Type I error rates has led to suggestions in the randomization test literature of mixing responsive and randomized components. In particular, Edgington (1975), Ferron and Ware (1994), and Koehler and Levin (1998) have suggested that researchers could initially gather data until stability is achieved and then make a random assignment and carry out a randomization test. Although this does allow for the initial data to be gathered in a response-guided fashion, the opportunity to respond to the data ends when the random assignment is made. If one randomly chose to intervene with the first participant after three more observations, the intervention has to occur after three more observations even if stability in the data has been lost (e.g., the third observation after assignment was an outlier). Furthermore, the intervention point for each of the other participants is set during the randomization process, which means that the researcher does not have the option of waiting until a treated behavior reaches criterion before starting treatment with the next participant. The researcher also would not be able to extend treatment phases to accommodate delayed, gradual, or small effects.

The review of the literature suggests that response-guided experimentation is valued, Type I error control is valued, and there is no method currently available that ensures Type I error control when analyzing data from response-guided multiple-baseline designs. Our purpose in this study was to provide such a method. The method occupies the middle ground between the visual analysis tradition and the randomization test tradition. From one perspective, it can be seen as a modification of visual analysis methods to incorporate a mechanism for controlling the Type I error rate. From the other perspective, it can be seen as a modification of randomization test methods that allows for more fully response-guided designs and allows, but does not require, the researcher to summarize the data with a test statistic. The latter point may be seen as advantageous by researchers who are concerned that a summary statistic may not fully summarize the changes that take place. For example, a difference in phase means does not fully capture the effect when there is a change in both level and variability, when there is a trend in baseline, or when the effect is delayed or occurs gradually.

Tests for Visual Analysts

The method we propose is based on separating the data analysis tasks from the other tasks involved in conducting a multiple-baseline study. The data analysis tasks could be shared by multiple researchers, as could other tasks. For descriptive simplicity, however, we will refer to only two researchers: a data analyst, who is responsible for all data analysis activities, and an interventionist, who is responsible for all other tasks. The interventionist will identify participants, plan and carry out the interventions, and make observations during baseline and treatment phases. The interventionist role differs in two ways from the traditional role of a researcher conducting a multiple-baseline study. First, each time a treatment is to begin, the interventionist randomly selects which participant (behavior, setting) will be treated. Second, the interventionist enlists the help of a data analyst who will be responsible for analyzing the data.

The data analyst conducts all data analysis tasks but is blind to which participant is selected for treatment at each designated intervention time. To keep the data analyst blind to which participants are and are not in treatment, it is critical that the data analyst does not make observations or have any direct contact with the participants. Rather, the interventionist must make all observations and then send the data to the data analyst. Note that the notion of a blind visual analyst has been presented elsewhere (Ferron & Foster-Johnson, 1998; Mawhinney & Austin, 1999). We are extending these ideas to multiple-baseline designs and integrating the work of the data analyst in a manner that allows one to calculate

Type I error rates for response-guided studies. The data analyst is responsible for plotting the data, deciding when data have stabilized, when the interventionist should intervene, and when sufficient data have been collected. Finally, the data analyst must make a data-based argument suggesting which participant (behavior, setting) was treated at each of the designated intervention points.

The method for controlling Type I errors in response-guided multiple-baseline studies is outlined in the Appendix. Note that the data analyst attempts to determine which participant (behavior, setting) was treated at each designated intervention time without having been told. If the intervention has no effect, the probability of determining the correct treatment assignment can be computed and the Type I error rate can be controlled. If the treatment has a large effect, the data analyst will be able to respond to the pronounced shifts in the data that result from the treatment and should be able to identify the treatment assignment. Finally, if the treatment effect is less pronounced, the data analyst has the flexibility to extend the phases until the pattern becomes clearer. Collecting this additional data should increase the odds of being able to correctly identify the treatment assignment. This method allows the researcher to accrue the benefits of response-guided experimentation, avoids the potential pitfalls of summarizing the data with a statistic, and allows the Type I error rate to be controlled.

Example

To more fully illustrate the method, we provide the following example. The interventionist has developed a social skills program designed to increase the percentage of behaviors that are socially appropriate for children with behavioral disorders. The interventionist selects four children to participate in the study. Figure 1 presents the hypothetical data for this study. After making each observation, the interventionist sends the data to the data analyst. The data analyst plots the data points as they are received. After graphing five observations for each child, the data analyst is not comfortable with the upward trend in Anne's data. The data analyst elects to wait, extending baseline beyond the agreed-on minimum of five observations. After the sixth observation is plotted, the data analyst is relatively comfortable with the stability of the data and tells the interventionist to start treatment with one of the participants.

The interventionist randomly selects one participant for treatment, begins treatment, and continues to gather and send data to the data analyst. The data analyst plots the next three observations. It appears that Pam is the one in treatment, but Emily's data are also trending upward. The analyst decides that the interventionist should gather more data. After plotting the fourth observation in this phase (the 10th overall), it is clearer to the data analyst that Pam is in treatment, whereas Emily is not. The data analyst then tells the interventionist to intervene with the next participant.

The interventionist randomly selects a participant, begins treatment, and continues to make observations, which are sent to the data analyst. After receiving and plotting three more observations, it is relatively clear to the data analyst that Anne has entered treatment, and thus the data analyst tells the interventionist to begin treatment with another participant.

The interventionist randomly selects a participant from the two who have not received treatment and begins treatment. The next two observations, which are made by the interventionist and sent to the data analyst, show a substantial change in Casey's behavior. The data analyst then tells the interventionist to administer the treatment to the last participant. The interventionist does so and continues to gather and send observations to the data analyst. After three more observations have been plotted, the data analyst believes sufficient data have been collected and tells the interventionist to stop gathering data.

At this point, the data analyst is relatively confident of the treatment order (Pam, Anne, Casey, Emily). Assuming for the moment that the treatment had an effect and that the data analyst correctly identified the order, the resulting graph is fairly convincing. A positive change in behavior was observed for each participant at the time treatment was implemented, and this change in behavior was maintained throughout the treatment phase. In addition, no appreciable change in behavior was observed for participants while they were in the baseline phase. Finally, the observed changes exhibit a temporal staggering, which matches the temporal staggering of the treatment and reduces the credibility of alternative explanations for the changes in behavior.

Because the data analyst correctly determined the treatment assignment, the *p* value can be computed by dividing one by the number of possible assignments. There are four participants to select from for the first intervention, three to select from for the second intervention, two to select from for the third intervention, and only one to select for the last intervention. Consequently the number of possible assignments is 24 (= 4! = 4 × 3 × 2 × 1). Thus the *p* value is or .0417, which is less than the conventional alpha of .05, and leads to the inference that the treatment had an effect.

Type I Error Control

The validity of this computed *p* value and the establishment of Type I error control are based on consideration of the probability that the data analyst would have correctly determined the treatment assignment if the treatment had not been effective. If the treatment had no effect, the same data would have occurred regardless of when the interventions were administered. Thus, all possible assignments would have led to the same data. Given the same data, it is assumed that the data analyst would make the same decisions about when to intervene and would come to the same conclusion about the treatment assignment. Thus, the same conclusion about the treatment assignment would be reached for any of the possible assignments. Because the assignment is made randomly, the probability of selecting the assignment that happens to correspond with the conclusion the data analyst will make is one divided by the number of possible assignments.

This argument parallels the one that is used any time a randomization test is used. For the *p*-value calculations to be accurate, only two assumptions are required. First, it is assumed that the same outcome data would lead the data analyst to the same decisions about when to intervene and the same conclusion about which participant was treated at each designated intervention time. This assumption holds as long as the decisions and conclusion made by the data analyst are based only on the outcome data, which, under the null hypothesis, are the same for all treatment assignments. This is why it is critical to provide the data analyst with the outcome data only and to keep the data analyst blind as to which participants are and are not in treatment. If the interventionist mentions who is in treatment, or if the data analyst makes some of the observations and thus discovers who is in treatment, the knowledge of who was in treatment could influence how the outcome data were viewed. In traditional randomization tests, selecting a test statistic prior to seeing the data is done for the same reason: It prevents knowledge of the treatment assignment from biasing how the data are analyzed.

The second assumption is that all possible assignments are equally likely. This assumption holds as long as the assignments are made randomly using a process that ensures equal probabilities. One such process would be to assign participants numbers. A random number stream could then be generated using software or by repeatedly rolling a fair die. The order for intervention could then be based on the order the participant numbers occurred in the random number stream. Note that random assignment is central to establishing equal probabilities. A researcher who systematically selects baselines for intervention cannot guarantee equal probabilities, and consequently the computed *p* value becomes questionable. Consider an interventionist who instead of randomly selecting the first participant for intervention tends to select the participant with the highest level of problematic behavior. If the data analyst is aware of this tendency, the probability of correctly determining the treatment order is increased and Type I error control is compromised.

Consider again the example shown in Figure 1. If the treatment had no effect, the data were not influenced by the treatment. Consequently, the data in Figure 1 would have arisen under any of the possible assignments. The shifts in behavior that are seen must be the result of factors other than the treatment; consequently, they would be there regardless of when the interventions occurred. We assume that the data analyst would arrive at the same conclusion that the treatment order was Pam, Anne, Casey, and Emily no matter what the actual assignment order was. This assumption will hold as long as the data analyst is making the conclusion based only on the data in Figure 1. If we assume that the assignment was made randomly, each of the 24 possible assignments had an equal chance of being made. The probability that the order Pam, Anne, Casey, and Emily was selected is $^1/_{24}$. Thus, there is a 1 in 24 chance of concluding that there was an effect if there was not.

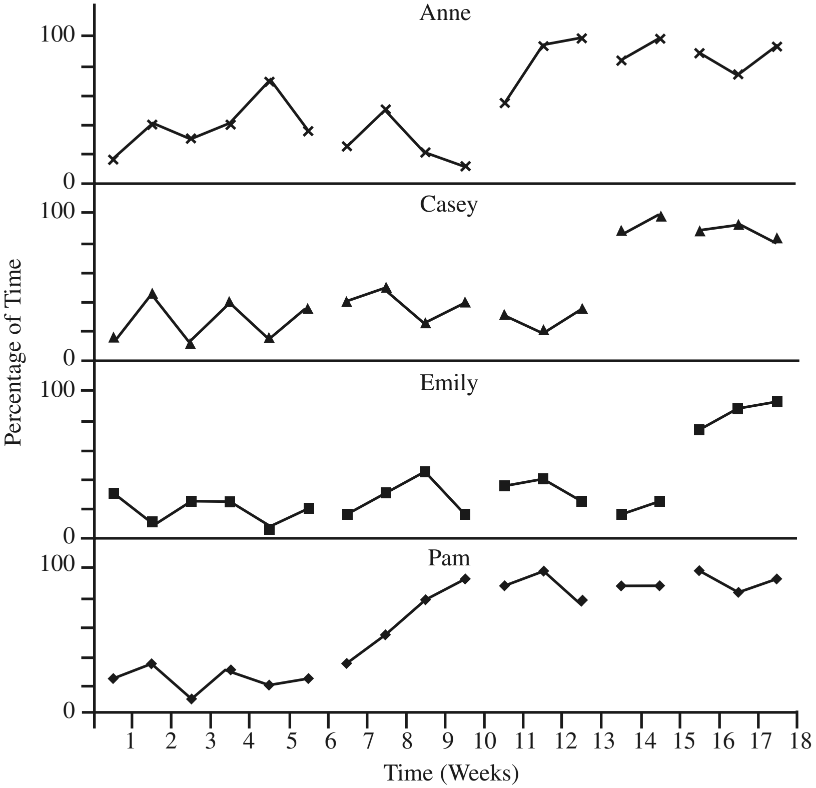

FIGURE 1. Graphical display of multiple-baseline data showing the percentage of time that the child demonstrates an appropriate social behavior. The graph does not explicitly show which participant enters treatment at which point. The intervention occurred between the 6th and 7th observation for one participant, between the 10th and 11th for another, between the 13th and 14th for another, and between 15th and 16th for the remaining participant.

If a researcher concludes only that there is a treatment effect when the probability is less than some specified level, say .05, the Type I error rate will be equal to or below the specified level. We have assumed the data analyst suggests a treatment order. If data analysts were allowed to not choose when they did not see an effect, the Type I error rates would be even lower.

Designs With More Than Four Baselines

The described visual test can be modified to make it compatible with a variety of multiple-baseline applications. As an example, consider a design involving more than four baselines. As the number of baselines increases, so does the number of possible assignments. With five baselines, there are 120 possible assignments; for six baselines, there are 720 possible assignments. As the number of possible assignments increases, the Type I error rate can be controlled to some smaller value, or the data analyst may be given more than one guess. As an example, consider a study with five baselines. Researchers wishing to control the Type I error rate to .05 could allow the analyst six attempts and maintain a Type I error rate of .05. If the data analyst guessed correctly on the sixth attempt, the *p* value would be $^6/_{120}$ or .05.

Designs With Three Baselines

Now consider a researcher who has only three participants in the study. Under the null hypothesis, the probability of guessing the first participant's intervention point correctly equals $^1/_3$, the probability of guessing the second participant's intervention point correctly equals $^1/_2$ conditional on the first guess being correct, and the probability of guessing the third participant's intervention point correctly equals 1 conditional on the first two being guessed correctly. The probability of making all three guesses correctly is $^1/_3 \times {}^1/_2 \times 1$, which equals .167. Consequently, it would be impossible to reject the null hypothesis of no treatment effect using a traditional alpha of .05.

The key to obtaining a smaller possible *p* value is to increase the number of possible intervention points for each participant. This can be accomplished by modifying the method so that the interventionist randomly selects without replacement from four possibilities (Participant 1, Participant 2, Participant 3, and no one) instead of from three possibilities (Participant 1, Participant 2, and Participant 3). This will lead to four possible intervention points for each individual instead of three. When there is no treatment effect, the probability of guessing the first participant's intervention point correctly equals $^1/_4$, the probability of guessing the second participant's intervention point correctly equals $^1/_3$ conditional on the first guess being correct, and the probability of guessing the third participant's intervention point correctly equals $^1/_2$ conditional on the first two being guessed correctly. Thus, the probability of making all three guesses correctly is .0417 when there is no treatment effect.

Discussion

Power

If one assumes the data analyst is blind to treatment assignments and that the assignments were made randomly, then exact *p* values can be calculated mathematically, and no experimental work is needed to show Type I error control. It is not possible, however, to derive Type II error rates mathematically. Experimental studies will need to be conducted to accrue power estimates for conditions that researchers are likely to encounter. One would anticipate that power would depend on a variety of factors, such as the size of the effect, type of the effect (e.g., immediate shift in level, delayed change in slope, change in level and slope, change in variability), number of baselines, series lengths, type of model underlying the error structure (e.g., first-order autoregressive, moving average), degree of dependence among the errors, method used to construct the graph, characteristics of the data analyst, and presence of historical, maturational, or testing effects. Power studies have started to emerge for more traditional forms of visual analysis (Fisch, 1998, 2001; Fisher et al., 2003; Matyas & Greenwood, 1990) and for more traditional randomization tests (Ferron & Sentovich, 2002). Hopefully, in future studies, researchers will estimate the power of these more traditional analyses for a wider range of conditions, as well as estimate the power of the visual tests for blind data analysts.

In thinking about estimating power, it is important to keep in mind that these tests were designed so that they could be used with response-guided experimentation, and response-guided experimentation allows phases to be extended when effects are less obvious. Consequently, to get a reasonable estimate of power, it will be important to simulate tasks that allow data analysts the opportunity to extend phases. This, of course, will add an extra layer of complexity in designing the power studies and will make it more difficult to directly compare the power of these tests with tests that assume that the phase lengths are established in advance. Nonetheless, this sort of research will be valuable in trying to get a sense of what kind and size of effects can be found using these tests. In addition, the sensitivity gains assumed to result from response-guided experimentation could be estimated, which would be useful in weighing the pros and cons of response-guided experimentation.

Agreement Among Analysts

Researchers should also look into the agreement among analysts. Studies have been conducted to examine agreement among analysts conducting traditional visual analyses (e.g., DeProspero & Cohen, 1979; Jones, Weinrott, & Vaught, 1978; Wampold & Furlong, 1981). The task of a blind data analyst, however, is different from the task of a traditional visual analyst. The blind data analyst uses the data to select the most plausible assignment from the set of possible assignments, whereas the traditional visual analyst uses the data and knowledge of the treatment assignment to make a judgment about whether the treatment was effective. Thus, additional studies will be needed to examine agreement of blind data analysts and the degree to which agreement is altered by training analysts or by using multiple analysts.

The Type I error rate is not dependent on the degree of agreement among analysts, thus, ensuring agreement will have no effect on the Type I error rate. It could, however, affect the Type II error rate. Consequently, attempts to increase agreement of analysts should be aimed at reducing the Type II error rate. The goal must be to find methods to help the less sensitive data analysts be more sensitive.

Training of Visual Analysts

The visual tests may be particularly useful for researchers who are less experienced in deciding how many observations need to be gathered to create a convincing argument that the treatment has had an effect. In addition, it may be possible to integrate these tests into a graphic analysis training program that generates emerging graphs such as the one proposed by Parsonson (2003). Research is needed to examine whether doing so would help visual analysts strengthen their skills in assessing stability, determining appropriate times to intervene, and accurately identifying treatment effects.

Applicability

Multiple-baseline designs are widely used in educational research. Part of this use stems from the accessibility of these designs to both practitioners and researchers who work in a variety of educational settings. Multiple-baseline designs provide a method of conducting action research for educational leaders (Glanz, 2003) and a method of conducting classroom research for teachers (Schloss & Smith, 1998). These designs have been useful to researchers in the areas of school psychology (Skinner, 2004), special education (Horner et al., 2005), counseling (Lundervold & Belwood, 2000), physical education (Ward & Barrett, 2002), and reading (McCormick, 1990).

The applicability of the method proposed here can be questioned, however, because it differs from the traditional multiple-baseline design in a couple of ways. First, a blind data analyst is used, which means that at least two researchers have to be involved in the research project. It appears that the inclusion of a blind data analyst would often be feasible because the time commitment of the data analyst is relatively small and the review of multiple-baseline applications found most studies involved more than one author. The second difference is the need for randomly assigning individuals (behaviors, settings) to the identified intervention times. This type of random assignment has been recommended elsewhere for multiple-baseline designs (Kazdin, 1982; Revusky, 1967; Wampold & Worsham, 1986) and appears possible in many educational contexts. Thus, it is anticipated that, even with the constraints introduced by the visual tests, these designs would be applicable in a wide range of educational settings.

Conclusion

The visual tests provide a method for controlling the Type I error rate when analyzing data from response-guided multiple-baseline designs. It is recognized that the addition of these tests provides little value for those who work in contexts where constant baselines can be obtained. In these situations, practically significant effects are obvious, and there is little question about Type I error control. Educational settings, however, typically lead to variable baseline data. In these circumstances, it is valuable to have an analysis method that controls the Type I error rate. In doing so, the visual tests provide a valid method for making treatment effect inferences. They do not, however, demonstrate that the treatment effect was of practical or clinical significance. This is a more complex assessment to make, but one that is aided by ruling out chance as a plausible explanation for the observed changes.

REFERENCES

Allison, D. B., Franklin, R. D., & Heshka, S. (1992). Reflections on visual inspection, response guided experimentation, and Type I error rate in single-case designs. *The Journal of Experimental Education, 61,* 45–51.

Armitage, P., McPherson, C. K., & Rowe, B. C. (1969). Repeated significance tests on accumulating data. *Journal of the Royal Statistical Society, Series A, 132,* 235–244.

Box, G. E. P., Jenkins, G. M., & Reinsel, G. C. (1994). *Time series analysis, forecasting, and control.* San Francisco: Holden-Day.

DeProspero, A., & Cohen, S. (1979). Inconsistent visual analysis of intrasubject data. *Journal of Applied Behavior Analysis, 12,* 513–519.

Edgington, E. S. (1975). Randomization tests for one-subject operant experiments. *The Journal of Psychology, 90,* 57–68.

Edgington, E. S. (1980). Validity of randomization tests for one-subject experiments. *Journal of Educational Statistics, 5,* 235–251.

Ferron, J., & Foster-Johnson, L. (1998). Analyzing single-case data with visually guided randomization tests. *Behavior Research Methods, Instruments, & Computers, 30,* 698–706.

Ferron, J., Foster-Johnson, L., & Kromrey, J. D. (2003). The functioning of single-case randomization tests with and without random assignment. *The Journal of Experimental Education, 71,* 267–288.

Ferron, J., & Sentovich, C. (2002). Statistical power of randomization tests used with multiple-baseline designs. *The Journal of Experimental Education, 70,* 165–178.

Ferron, J., & Ware, W. (1994). Using randomization tests with responsive single-case designs. *Behavior Research and Therapy, 32,* 787–791.

Fisch, G. S. (1998). Visual inspection of data revisited: Do the eyes still have it? *The Behavior Analyst, 21,* 111–123.

Fisch, G. S. (2001). Evaluating data from behavioral analysis: Visual inspection or statistical models? *Behavioural Processes, 54,* 137–154.

Fisher, W. W., Kelley, M. E., & Lomas, J. E. (2003). Visual aids and structured criteria for improving visual inspection and interpretation of single-case designs. *Journal of Applied Behavior Analysis, 36,* 387–406.

Glanz, J. (2003). *Action research: An educational leader's guide to school improvement* (2nd ed.). Norwood, MA: Christopher-Gordon.

Greenwood, K. M., & Matyas, T. A. (1990). Problems with the application of interrupted time series analysis for brief single-subject data. *Behavioral Assessment, 12,* 355–370.

Hagopian, L. P., Fisher, W. W., Thompson, R. H., Owen-DeSchryver, J., Iwata, B. A., & Wacker, D. P. (1997). Toward the development of structured criteria for interpretation of functional analysis data. *Journal of Applied Behavior Analysis, 30,* 313–326.

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

Huitema, B. E. (1985). Autocorrelation in applied behavior analysis: A myth. *Behavioral Assessment, 7,* 107–118.

Huitema, B. E., & McKean, J. W. (1998). Irrelevant autocorrelation in least-squares intervention models. *Psychological Methods, 3,* 104–116.

Huitema, B. E., & McKean, J. W. (2000). A simple and powerful test for autocorrelated errors in OLS intervention models. *Psychological Reports, 87,* 3–20.

Jones, R. R., Weinrott, M., & Vaught, R. S. (1978). Effects of serial dependency on the agreement between visual and statistical inference. *Journal of Applied Behavior Analysis, 11,* 277–283.

Kazdin, A. E. (1982). *Single-case research designs.* New York: Oxford University Press.

Koehler, M., & Levin, J. (1998). Regulated randomization: A potentially sharper analytical tool for the multiple-baseline design. *Psychological Methods, 3,* 206–217.

Kratochwill, T., Alden, K., Demuth, D., Dawson, D., Panicucci, C., Arntson, P., et al. (1974). A further consideration in the application of an analysis-of-variance model for the intrasubject replication design. *Journal of Applied Behavior Analysis, 7,* 629–633.

Lundervold, D. A., & Belwood, M. F. (2000). The best kept secret in counseling: Single-case (N = 1) experimental designs. *Journal of Counseling and Development, 78,* 92–102.

Matyas, T. A., & Greenwood, K. M. (1990). Visual analysis of single-case time series: Effects of variability, serial dependence, and magnitude of intervention effects. *Journal of Applied Behavior Analysis, 23,* 341–351.

Matyas, T. A., & Greenwood, K. M. (1997). Serial dependency in single-case time series. In R. D. Franklin, D. B. Allison, & B. S. Gorman (Eds.), *Design and analysis of single-case research* (pp. 215–243). Mahwah, NJ: Erlbaum.

Mawhinney, T. C., & Austin, J. (1999). Speed and accuracy of data analysts' behavior using methods of equal interval graphic data charts, standard celeration charts, and statistical control charts. *Journal of Organizational Behavior Management, 18,* 5–45.

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

McKnight, S. D., McKean, J. W., & Huitema, B. E. (2000). A double bootstrap method to analyze linear models with autoregressive error terms. *Psychological Methods, 5,* 87–101.

Padia, W. L. (1976). The consequences of model misidentification in the interrupted time-series experiment (Doctoral Dissertation, University of Colorado, 1975). *Dissertation Abstracts International, 36,* 4875A.

Parsonson, B. S. (2003). Visual analysis of graphs: Seeing is believing. In K. F. Budd & T. Stokes (Eds.), *A small matter of proof: The legacy of Donald Baer* (pp. 35–51). Reno, NV: Context Press.

Parsonson, B. S., & Baer, D. M. (1992). The visual analysis of data, and current research into the stimuli controlling it. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case research design and analysis: New directions for psychology and education* (pp. 15–40). Hillsdale, NJ: Erlbaum.

Revusky, S. H. (1967). Some statistical treatments compatible with individual organism methodology. *Journal of Experimental Analysis of Behavior, 10,* 319–330.

Robinson, D. H., & Levin, J. R. (1997). Reflections on statistical and substantive significance, with a slice of replication. *Educational Researcher, 26,* 21–26.

Schloss, P. J., & Smith, M. A. (1998). *Applied behavior analysis in the classroom* (2nd ed.). Boston: Allyn and Bacon.

Skinner, C. H. (2004). Single-subject designs: Procedures that allow school psychologists to contribute to the intervention evaluation and validation process. *Journal of Applied School Psychology, 20,* 1–10.

Stocks, J. T., & Williams, M. (1995). Evaluation of single subject data using statistical hypothesis tests versus visual inspection of charts with and without celeration lines. *Journal of Social Research, 20,* 105–126.

Todman, J., & Dugard, P. (1999). Accessible randomization tests for single-case and small-n experimental designs in AAC research. *Augmentative and Alternative Communication, 15,* 69–82.

Velicer, W. F., & Harrop, J. (1983). The reliability and accuracy of time series model identification. *Educational Review, 7,* 551–560.

Wampold, B. E., & Furlong, M. G. (1981). The heuristics of visual inspection. *Behavioral Assessment, 3,* 79–92.

Wampold, B., & Worsham, N. (1986). Randomization tests for multiple-baseline designs. *Behavioral Assessment, 8,* 135–143.

Ward, P., & Barrett, T. (2002). A review of behavior analysis research in physical education. *Journal of Teaching in Physical Education, 21,* 242–266.

Westra, D. P. (1979). Testing interventions in the interrupted time series quasi-experiment: The reliability of Box-Jenkins noise model specification with short series (Doctoral dissertation, University of South Dakota, 1978). *Dissertation Abstracts International, 39,* 5621B.

APPENDIX

Steps for Method of Conducting Response-Guided Multiple-Baseline Study in Which Type I Error Rate Is Controlled

1. The interventionist plans study, recruits participants, and enlists help of a data analyst.
2. The interventionist and data analyst discuss the details of the study. They discuss the minimum number of points that will be gathered during a phase, how decisions about stability will be made, whether criteria need to be met by participants in treatment before other participants enter treatment, and what magnitudes and types of effects can be anticipated.
3. The interventionist starts the study and collects baseline data for each participant. After each observation, the data is sent to the data analyst.
4. The data analyst plots and analyzes the data received from the interventionist. After the minimum number of observations has been collected and the data are judged to be stable, the data analyst tells the interventionist that treatment should be started with one of the participants.
5. The interventionist randomly selects a participant for intervention. The interventionist then begins the treatment and continues to make observations that are sent to the data analyst. Note that the interventionist does not tell the data analyst which participant is being treated.
6. The data analyst continues to plot and analyze the data. When the data analyst has enough data to make a reasonable argument as to which participant is being treated, the data analyst tells the interventionist that treatment should be started with the next participant.
7. Steps 5 and 6 are repeated until all participants have been treated.
8. The data analyst signals the end of the study when sufficient data have been collected for each baseline and treatment phase.
9. The data analyst constructs an argument based on the data, indicating which participant received the intervention at each of the designated intervention points.
10. The interventionist then discloses the actual assignment. If the data analyst identified the assignment correctly, a *p* value is computed by dividing one by the number of possible assignments. If the data analyst failed to identify the assignment correctly, the null hypothesis is not rejected.

Note. The steps were written assuming a multiple-baseline across-participants design. For multiple baselines across behaviors or settings, the steps would be parallel and references to participants would be replaced with references to behaviors or settings.