Chapter 4
Evaluation and Analysis of Data Generated from Single Subject Designs

The methodological sophistication of single subject designs has been discussed since their introduction by R.A. Fisher in 1945 [1]. This chapter will cover the major approaches used in evaluating and analyzing data from single subject designs, especially as applied to patient or clinical care, along with outcome research assessing the therapeutic effect of the intervention (i.e., evidence based practice) [2]. Claude Bernard, the father of experimental medicine, provided the broad foundation for the application of the experimental method to practice-based research in medicine [3]. Furthermore, he proposed that the use of statistical techniques to interpret data should be cautioned. He held that statistics can only lead to probabilistic estimates, which in his time were contrary to the prevailing philosophy that scientific laws should possess deterministic certainty. Bernard also postulated that certainty could ultimately be achieved with investigator insight and the application of rigorous experimental controls. Although the use of statistics is commonplace and essential in contemporary research, Bernard's wisdom regarding the importance of conducting a sound study should not be ignored. Applying statistics to poorly conceived and designed studies will not save or increase the validity of such studies; rather, it might lead to some long lasting misconceptions that could negatively impact the welfare of patients. The coverage in this chapter will focus on evaluation of the data that are generated by single subject research and techniques for displaying and analyzing data collected through single subject studies.

Experimental Control and the Single Subject Design

Barlow, Nock, and Hersen [4] argued that in order to establish clinical science, it is important to determine the sources of variability in individuals. Variability occurs within an individual (intra) and between individuals (inter). Determining the sources of variability allows the researcher to reduce measurement error. In turn, this approach allows for the establishment of a causal relationship between the independent (i.e., intervention) and dependent (i.e., outcome) variables, thus enhancing the internal validity (See Chapter 3).

J.E. Janosky et al., *Single Subject Designs in Biomedicine*,
DOI 10.1007/978-90-481-2444-2_4, © Springer Science+Business Media B.V. 2009

It is important that data collected by the researcher be as free as possible from alternative explanations or hypotheses thus affording the researcher the ability to state emphatically that the change in the dependent variable is due to the independent variable and not to some other variable. In other words, the researcher should be able to conclude that the study is internally valid [5]. It is critical that when conducting an experiment (defined as any study where the researcher has control over the presentation or withdrawal of the intervention) special care is taken. Unfortunately, it is rarely the case that the researcher can control or rule out all "other" variables; therefore, the data are not entirely free from alternative explanations or hypotheses. Consequently, a sound study is one in which the alternative explanations or the threats to internal validity are not plausible [2].

The essence of any study, including single subject studies, is the utilization of proper controls [6]. Rigorous controls minimize the role of error. Within the context of a study, control refers to the ability of the researcher to influence or change (i.e., manipulate) the variables in a study. However, before one can apply the experimental controls to a study, one must first identify the possible sources of error (i.e., extraneous variables) in the methodology of the investigation. In other words, one needs to evaluate the methodology used for the study, as the methodology dictates the conditions for data generation. If there are limitations or flaws in the methodology, there are likely to be limitations or flaws in the data that will likely impact study conclusions. It is also important to note the unique features of single subject research and their relationship to control. It is common practice in this type of research to repeatedly test one or a few patients over an extended period of time with multiple points of evaluation (i.e., outcome measurement). Single subject research differs from the more traditional between subjects large N research where randomization of patients to interventions is used to control for individual differences. Rather, in single subject research, control is achieved for individual differences through each patient being used as his or her own control (intra-subject). Specifically, the researcher is comparing each patient's outcome measure during baseline (pre-intervention) and intervention. Although this is sound methodology for controlling individual differences, one of the negative consequences of this approach is that there may be transfer, or carry-over effects, from repeated treatments or interventions. Therefore, the unique properties of single subject designs need to be recognized when attempting to control for extraneous variables.

An extraneous variable is defined as any variable which may impact the target outcome, but it is not the intervention or treatment (i.e., independent variable) [7]. Extraneous variables threaten the internal validity of a study if the following conditions exist: First, the extraneous variable is systematically related to the intervention or treatment, or the variables co-vary; and second, the extraneous variable is systematically related to the outcome. Uncontrolled variables that co-vary with the intervention and influence the outcome produce a confounded study. In this case, the intervention is not solely responsible for study effects, as multiple explanations exist. For example, a major assumption in the popular A-B-A design pertains to constancy of conditions, in which the only change from the baseline to the treatment, or treatment to baseline, is the presentation or removal of the intervention. The study

is of limited pragmatic value if this assumption is violated and the introduction of the independent variable is correlated with the introduction of an extraneous variable, which in turn influences the dependent variable. If there is covariation within phases of a single subject study, then it is possible the study is confounded and that the researcher influenced the outcome. For example covariation could occur if one researcher collects the data during baseline conditions and another researcher collects the data during the treatment condition. The results of such a study would be highly suspect, given the lack of empirical evidence that the extraneous variable does not influence the outcome. Finally, it should be noted that if there is no systematic relationship between the extraneous variable and intervention, then there is no concern as to whether the extraneous variable influenced the outcome. Nonetheless, it is still important to control for extraneous variables, since these variables can reduce the sensitivity of the intervention, therefore contributing to the random error or noise in the study.

Techniques of Control

There are a number of general control techniques that can be used to eliminate or reduce the influence of extraneous variables in a study [7]. These techniques will be listed and described below. It is important to keep in mind that the techniques are listed in terms of their power or ability to control extraneous variables. Also, It would be useful to employ these techniques as a checklist for deciding what controls to use in a single subject study.

Elimination

If an extraneous variable exists in the study and it can be identified, the first step would be to determine if it can be removed from the study. If the extraneous variable can be removed, then it will not confound the results. Unfortunately, this technique cannot be used very often because most extraneous variables are an integral part of the study setting. For example, it would be impossible to eliminate the medical histories of the patients. If it is unlikely that an extraneous variable can be eliminated, there may be potential extraneous variables that can be reduced to levels where it is highly unlikely to have any effects. For instance, the ambient noise levels can be reduced in a research setting, eliminating this variable as an extraneous variable impacting the setting.

Constancy

If the extraneous variable cannot be eliminated, an attempt should be made to hold constant the extraneous variables. Constancy is achieved when the identified extraneous variable occurs in all of the phases or conditions of the study with the same quantitative properties. Many potential extraneous variables can be controlled using this technique. For example, it is important to make sure the study is conducted in

the same setting for each patient, testing occurs at approximately the same time of the day, instructions are standardized, and testing is completed by the same recorders or evaluators.

It is also important to recognize that constancy is a very useful principle to apply even before the actual start of the study. It is common for some patients to exhibit physiological (e.g., increases in blood pressure) and psychological (e.g., increases in anxiety) anticipatory signs before entering the actual study environment. In essence, simply waiting to be tested may ultimately reduce the sensitivity of the treatment or intervention leading to Type II errors (i.e., the failure to detect an actual effect). Therefore, constancy can be a valuable technique to use for the entire single subject environment. Although constancy is an excellent control technique that can be used to manage extraneous variables, it is not foolproof. For example, even though all patients are tested at the same time of the day, it does not follow that all will respond in the same manner to the same testing time. Additionally, as discussed in Chapter 3, attempts to control for confounding variables are ineffective to the extent that they suffer from poor measurement reliability.

Balancing

If extraneous variables are not amenable to the technique of constancy, it may be possible to use balancing. In the case of balancing, the extraneous variable is equalized across the conditions or phases of the study. It is important to distinguish between balancing and constancy. For example, in an A-B design, if constancy is being used to control for the testing environment, all patients would be tested in this same testing environment. On the other hand, due to practical necessity, the researcher may be required to test in more than one setting. In this case, it would be important to balance patients across the research settings. This could be accomplished by randomly assigning patients to treatment settings with the restriction that an equal number be placed in each treatment environment. Not only have extraneous variables been controlled using this technique, but the effects of the extraneous variable can be assessed by comparing the target variable across the settings. It is important to note that balancing and constancy achieve the same objective of controlling for the extraneous variable, but constancy is a more powerful technique. In comparison to balancing, constancy results in little if any variance in the extraneous variable across the phases of the study. When error variance or noise in the study is reduced, the accuracy and validity of the results increase.

Counterbalancing

Counterbalancing is more likely to be used in single subject designs than balancing. In contrast to the latter technique, counterbalancing is used when each patient serves in two or more treatments or conditions (i.e., a repeated measures or within-subjects design). Counterbalancing is frequently used when the researcher

suspects carry-over or order effects will occur across the treatments. This affords the researcher the ability to assess the effects of the treatments, since the treatments are not contaminated by the order in which they are presented. For example, the physician or researcher may be interested in testing the therapeutic effectiveness of three different medication dosage levels. In this A-B-A-C-A-D design, the baseline (A) is established and reestablished after each level of the treatment is administered (B, C, and D). Counterbalancing can be achieved by first determining the number of permutations or orders among the treatments. In this case, we have three treatments. Using the expression, $n! = n(n–1)\ (n–2)$ until $(n–(n+1))$, where n equals the number of treatments, the number of possible orders is six. The six orders are B-C-D, C-D-B, D-B-C, D-C-B, B-D-C, and C-B-D. Note that each treatment precedes and follows every other treatment an equal number of times. Unfortunately, a minimum of six subjects would be needed to use this form of counterbalancing (called complete counterbalancing). Patients would be randomly assigned to the orders or the sequences of the treatments. If it is not practical to use six patients, then the number of patients required may be reduced by randomly selecting a subset of orders (called incomplete counterbalancing). Since not all permutations are represented in incomplete counterbalanced designs, as compared to complete counterbalanced designs, the strength of the incomplete counterbalanced design is less than that of the complete design. The major assumption of counterbalancing is that the effects of order will balance out; for example, the effects of B on C will equal the effects of C on B (symmetrical transfer). Unfortunately, it is possible to find asymmetrical transfer, in which transfer differs depending on the order (See McGuigan [7], for a more in-depth discussion of counterbalancing).

Randomization

Randomization has been mentioned in the previous discussion concerning techniques of control. However, randomization is also a major control technique. Randomization is a first line means of achieving control, as each element in a set has an equal chance of being selected. It is particularly appropriate when the other techniques cannot be used or when the researcher suspects the existence of extraneous variables, but is not able to identify them. In the long run, randomization is assumed to "balance out" the effects of these unknown variables. Randomization will be discussed in more detail later in this Chapter.

Interventions (Independent Variables)

The discussion concerning study control detailed the identification and control over extraneous variables. The implicit assumption was made that the independent variable was present in the form that was intended and that was accurate. The researcher needs to demonstrate that the intended intervention is the independent variable in the study, or that the study possesses treatment integrity or fidelity [8–9]. Treatment integrity also includes treatment differentiation. Treatment differentiation refers

to studies where the goal is to compare the effects or outcomes of two or more treatments. It is important to establish that the treatments are sufficiently different such that the comparison is legitimate. That is, the researcher can safely conclude that if no differences were found between the treatments, failure to establish treatment differentiation was not responsible.

Gresham [8] has described in some depth the role of treatment integrity, also known as treatment fidelity, and its relationship to internal validity. In essence, if the intervention is not presented accurately and consistently and effects are found, the researcher may falsely conclude that the intended intervention is responsible for the outcome (i.e., a Type I error). Also, failure to present the intended independent variable may lead to no outcome effects, and the researcher may falsely conclude that the independent variable was not effective when it was effective (i.e., a Type II error). Overall, failure to establish treatment integrity weakens the internal validity of the study.

Treatment integrity appears to be a trivial issue for single subject researchers. Based on previous literature reviews, Gresham [8] concluded that the majority of researchers did not attempt to establish treatment integrity. It is important to emphasize that it is difficult to rule out alternative explanations if the physician or researcher fails to establish treatment integrity, or treatment differentiation. Treatment integrity or differentiation may be particularly important to establish when the treatment is complex. The treatment or independent variable must be reliable, valid, and accurate. It is therefore critical that care is taken in operationalizing the independent variable; that is, converting the conceptual definition of the independent variable into an observable, measurable, and verifiable definition that is accurate and precise. In essence, there is a high correspondence between conceptual definition and the measured definitions [8, 10, 11]. Gresham [8] describes some methods for assessing treatment integrity, including direct assessment (e.g., systematic observation) and indirect assessment (e.g., rating scales, interviews, self-monitoring, and self-reports). The type of research and nature of the independent variable guides the researcher in selecting which methods are most appropriate for patient or clinical care research. Finally, Gresham [8] recommends the use of the dependability index in providing estimates of reliability and validity in single subject research [12, 13].

Outcomes (Dependent Variables)

A corollary to treatment integrity is the selection and measurement of the dependent or outcome variable, a topic that is particularly important in research dealing with patient care. Measurements can be obtained through direct observation, automated recordings, rating scales, and checklists, for example. As has been emphasized in the literature, the selection of the dependent variable should be based on its practical, social, or medical significance. The outcome needs to be directly relevant and beneficial to the patient's welfare, which is interpreted as such by the patient [14]

and the community [4, 14–16]. Furthermore, the measurement of the outcome needs to meet the requirements of reliability, validity, and accuracy [2, 15–17]. Reliability refers to a measure of consistency or repeatability of the outcome variable. Validity refers to the extent to which the target outcome is measured directly, which is the focus of the study. Accuracy refers to the extent to which the measured observation matches the true state of the event. For example, does a measure of blood pressure produce similar results each time it is measured under the same conditions (reliability)? Also, is it measuring blood pressure as it purports (validity) and is the actual value obtained with the measuring instrument the true state of affairs (accuracy)? It is important to recognize that all of these requirements must be established before meaningful conclusions can be determined concerning the influence of the independent variable, or intervention, on the outcome measure [15].

Numerous methods have been proposed for establishing accuracy, reliability, and validity (see Cooper, Heron, and Heward [15] for a rendition on measurement in single subject research). In the case of validity, there are direct and indirect measures [15]. Direct measures are a reflection of the phenomenon under investigation. Indirect measures occur when the actual measurement is not directly related to the phenomenon, and therefore requires more of an inference on the part of the researcher. It is best to keep in mind that direct and indirect measures are relative; for example, the arm cuff (i.e., the sphygomomanometer) would be viewed as more of a direct measure of blood pressure, whereas self report would be viewed as more of an indirect measure. Direct measures typically show higher validity than indirect measures. However, sometimes direct measures are not available and the researcher must resort to indirect measures. For example, if the researcher is interested in the mental status of the patient, an indirect measure may be the best solution. Regardless of type of measurement, it is important that validity be established. The establishment requires that the researcher provide evidence that the phenomenon under investigation is in fact being measured.

With behavioral measurement and subjective measurements, and because human error is one of the biggest threats to reliability and accuracy, it is common practice to use inter-observer agreement (IOA). IOA refers to the extent to which two or more independent observers report the same values in assessing reliability and accuracy of the measurements [2, 15–17]. Although percentage of agreement is the most common technique for measuring IOA, there are many other techniques as well [15]. Furthermore, it is important to recognize that considerable time and effort must be attached to the selection and training of the observers in order to avoid or reduce bias or artifacts [2, 15, 17]. For example, bias can occur in the data because of observer drift (the observer changes the definition of what is to be observed during the course of the study), observer reactivity (the observer is sensitive to the notion that her/his observations are being evaluated by someone else), and observer expectations (the observer is aware of the predictions or hypotheses of the study). Also, ultimately, the researcher must decide on a criterion for determining whether the data are reliable and accurate. The standard acceptance level for a numerical cut-off for quantitative measures of reliability in the literature, is 0.80. However, Kazdin [2] and Cooper et al. [15] have argued that it is not wise to set a rigid

criterion because the criterion of acceptance depends on the nature and complexity of the research. Finally, Primavera, Allison, and Alfonso [17] noted that the failure to establish reliability is widespread in single subject research. It may appear obvious to the researcher that the dependent measure is reliable; however, without some assessment of its reliability, it would be difficult for the researcher to claim, for example, that the failure for finding a relationship between the intervention and outcome is due to the ineffectiveness of the treatment.

Response Guided Studies

A section of this Chapter is devoted to response guided study because of its central role in single subject research and because of the debate concerning internal validity. A tactic integral to single subject research, especially research with practical or clinical significance, is termed response guided experimentation [1, 18]. This strategy refers to the common practice in single subject research where the researcher or physician makes decisions during data collection regarding the length of the baseline, along with the timing to present and withdraw the treatment [2]. The goal of this strategy is to change the baseline and treatment variables in such a way as to maximize the effectiveness (or lack thereof) of the treatment [2]. In other words, rather than having a structured research plan for the conduct of the study, the researcher changes the phases of the study based on the patient's responses. Although Kazdin [2] has suggested some tips for making these decisions, such as examining the trends and variability in the data, there are no well established decision rules for determining these changes; therefore, it is largely based on the experience and assessment of the researcher [2]. Edgington [1] has argued that this approach is more art than science. He points out the potential flaws in this approach, including issues that limit the quality of the data, are based upon the competence of the researcher, the lack of objectivity for the approach, and perhaps most importantly, the possibility that the researcher and treatment are confounded. The confound is especially critical because it is difficult to ascertain whether the changes, if any, were due to the treatment or due to the researcher effects (e.g., expectancies).

In support of response guided experimentation, Barlow and Hersen [19], Kazdin [2], Krishef [20], and Barlow et al. [4]. have strongly recommended the use of this approach. They stress the clinical significance of determining the source of variability in individual patients and the compatibility with standard clinical practice. Single subject research is ideally suited for the physician or researcher. One can observe the variability of the individual patients during baseline (A) and treatment (B), speculate or hypothesize about the sources (i.e., the causes), and immediately adjust the design, so as to test these hypotheses. Consequently, the welfare of the patient is likely to be enhanced. In order for this approach to be successful, it is essential that repeated testing be employed with the requirement that the physician or researcher have the ability to change the research design as needed. It is apparent that single subject research is ideally suited for meeting these requirements. It is also important

to recognize that with these essential features, single subject research is of added value to the physician.

Barlow et al. [4] have suggested three ways in which these improvised or rapidly alternating single subject designs can be used in determining the sources of variability, which can possibly improve the internal validity. These include cases in which the patient fails to improve with a given treatment, the patient improves spontaneously (i.e., placebo effects or improvement occurs in the absence of the treatment), or the patient's outcome measure exhibits cyclic patterns across and/or within phases. In each case, a common tactic is to change the design to see if the causes of the variation can be identified. Finally, Barlow et al. [4] indicated that in many clinical cases, the sources of variability may be difficult to identify, called hidden sources, and may involve a multiplicity of variables, as well as interaction effects. Therefore, it behooves physicians to apply their experience and evaluative skills before deciding on the causes of the outcome. This strategy was applied with remarkable success by the father of experimental medicine, Claude Barnard [3]. Overall, as Barlow and Hersen [20] and Houle [21] have stated, the criterion of evaluation is that the study must meet the requirements of internal validity, and the results must be therapeutically meaningful to the patient.

Statistical Analysis of Data Collected Using Single Subject Methodology

In a research based monograph by physicians and psychologists [22] considerable research is presented suggesting that many physicians and a significant portion of patients exhibit statistical illiteracy; in essence, statistical illiteracy is the failure to accurately interpret the numbers when assessing the risks and benefits of foregoing or undergoing treatment. Statistical illiteracy may not necessarily be a failure of understanding the numbers per se, but more a result of cognitive biases, physician-patient relationships, and conflicts of interest [22]. Regardless of the cause, statistical illiteracy or the failure to properly interpret health related statistics can lead to dire consequences for the patient.

The problem of statistical illiteracy is not unique to health providers. Gigerenzer et al. [22] have coined the expression "collective statistical illiteracy" reflecting their view that statistical illiteracy is a widespread societal problem. Consistent with this notion, Monahan [2] points out that statistical illiteracy is common among judges and juries. In fact, American tort law still to some extent encourages the use of the antiquated legal standard of care in which physicians must demonstrate that the prescribed treatment was based on current standard of care rather than evidence based practice. For these reasons, as concluded by Gigerenzer et al. [22], it behooves practitioners to become more statistically literate in order to function competently as professionals. Provided here are some statistical procedures, both descriptive and inferential, that are applicable to evaluating data collected through single subject design methodology.

Graphical Display of Data and Visual Analysis

Visual analysis (also called the interocular test, eyeballing the data, criterion by inspection, or visual inspection) refers to the interpretation of data that have been plotted on a graph [15] without any additional statistical analyses. Despite the debates concerning validity, visual analysis is still commonly used for evaluating data generated from single subject designs. In order to interpret the finding of a study using visual analysis, it is critical that the data be properly graphed. There are numerous sources on appropriate procedures for displaying data from single subject designs [4, 15, 20, 23–25]. This section will focus on presenting and interpreting data from a graph using the methods that have been typically used in single subject research.

There are a number of benefits to using graphs. Houle [21] noted that "There is no replacement for the information provided by graphing the outcome variable as it varies over time" (p. 272). Houle [21] and others [15, 16] have stressed the importance of graphs in showing the variability in the data, as well as communicating the results to researchers and patients. Cooper et al. [15] and Parsonson and Baer [24] described the benefits of providing the researcher with an ongoing visual record of the progress of the study, changing the baseline (A) and/or intervention (B) based on the graphed data (i.e., response guided experimentation), providing an independent and more conservative approach (by noting only strong effects in the data and ignoring weak effects that may be statistically significant but not clinically significant), and providing the patient with an ongoing record of progress in the study.

If any benefit is to be derived from visual analysis, it is critical that standardized procedures be used for displaying the data. One of the most important but simple rule to follow is that the data points and data paths need to be accurately plotted [15]. Although software programs (See Carr and Burkholder [25] and Silvestri [48], How to make a graph using Microsoft Excel. Unpublished manuscript) are commonly used to construct graphs, it remains important to be able to graph relationships by hand, especially in response guided studies. The physician or researcher needs to have an ongoing visual record of patient outcomes to treatment, so that the treatment can be altered if necessary. Figure 4.1 depicts the results of a single subject study. The purpose of this study was to examine the effects of a medication on systolic blood pressure. Although some of these data were taken from an actual patient, for purposes of illustration, some data points were changed and additional data points were added. Copper et al. [15] strongly recommend that before attempting to understand the relationships among the data through visual analysis, it is very important to understand the basic features of the construction of the graph (e.g., the labeling and scaling of the axes, examination of the data points, and their linkage). Without a careful examination of the basic features of the graph, the interpretation of the relationships is more susceptible to human error [11]. Komaki, Coombs, Redding, and Schepman [26] recommend using a set of criteria called OCT for evaluating data from single subject designs. First, the researcher should examine the overlap (O) in data points between phases, then examine the measure of central tendency (C) for each phase, and finally look for subsequent trends (T).

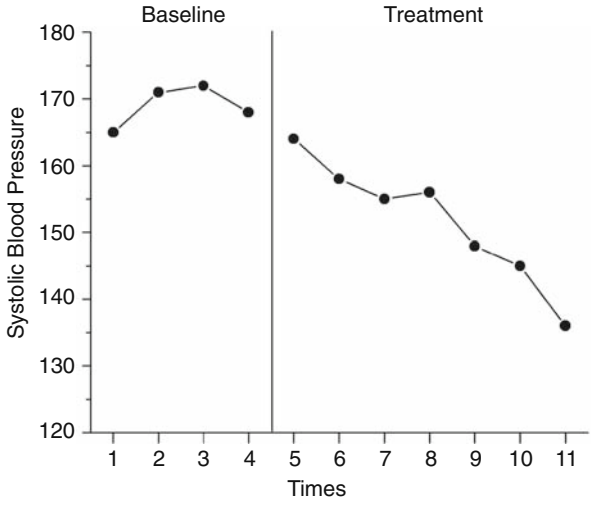

Fig. 4.1 An example of an A-B design

More comprehensively, Cooper et al. [15] stressed the importance of first applying visual analysis within phases, including the number of data points, variability, level, and trend. Next, examine the data between conditions using the same criteria (i.e., number of data points, variability, and trend). Finally, Shadish, Cook, and Campbell et al. [5] suggested that the researcher should examine whether the treatment effects will decay over time and whether this decay is immediate or delayed. Implementing this recommendation may require some follow-up tests after the initial stages of the study have been completed.

Figure 4.1 displays an example of an A-B design. First, baseline (A) measurements were obtained without any medications or interventions. The data in the baseline phase (A) appear relatively stable. Next, the patient received a medication treatment (B) that was intended to lower systolic blood pressure. Through visual inspection, the relationship seems apparent across conditions, as the medication appears effective in reducing systolic blood pressure. Relative to the baseline (A) levels, it is clear that systolic blood pressure lowered when the treatment was applied. It is important to note that the relationship between the treatment and systolic blood pressure might have shown further strengthening if an A-B-A design was implemented. Specifically, one of the strong features of the A-B-A design is that if the level of the outcome returns to baseline levels in the second baseline phase, the causal interpretation of the relationship is enhanced. Some descriptive statistics (i.e., measures of central tendency and variability) can be applied to these data because of the consistent variability within and between phases and the lack of any trend in the data. The best procedure [23] is to superimpose these measures on the plotted time series data. In this case, medians, a measure of central tendency reflecting the middle most score as represented by a continuous line in the graph, along with range lines, a measure of variability reflecting the low score and high score represented

by dashed lines, could be reported. For the evaluation *via* a clinical criterion, the overall evaluation might be driven by the meet/no meet level of sustained systolic blood pressure reading (e.g., 110 for systolic blood pressure). Means could also be used to represent these data, but "real data" from single subject designs are likely to include outliers or extreme scores, and medians are less influenced by these scores than means. A final concern with these data is the possibility that they are autocorrelated, a topic to be discussed later in this Chapter.

Unfortunately, in the actual conduct of research, interpretations are not as straightforward, as it is rare to find data demonstrating major effects with little variability or trends in the data. Figure 4.2 provides a more "realistic" view of data generated from a single subject design. These data are simulated for illustration. The major difference between Figs. 4.1 and 4.2 is that there is more variability in all phases and noticeable trends in the latter two phases in Fig. 4.2. Showing medians as points of comparison across the phases would not be meaningful because of trends in the data. In this situation it is advantageous to use the split-middle method [15, 20, 23, 27] to reflect trends in the data. In Fig. 4.2, the split middle method was used to create the line that is superimposed over the data points for each phase of the study. The line for each phase is calculated by dividing the data points into halves for each phase, then locating the median time value and median blood pressure measure for each half, plotting the coordinates for each half, and finally drawing a line connecting the two coordinates. As presented in Fig. 4.2, dividing the data points into

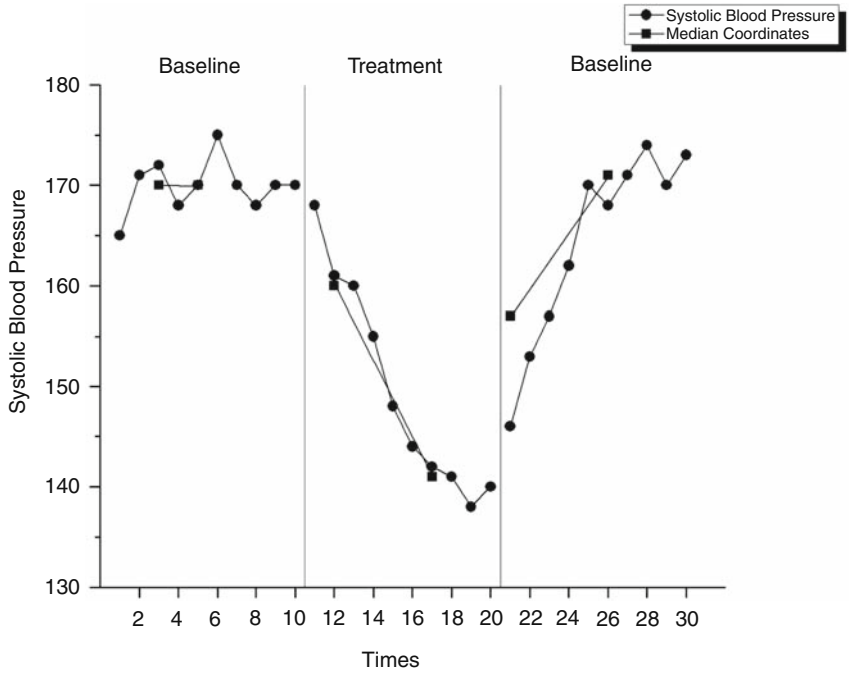

Fig. 4.2 Illustration of an A-B-A design

halves within each phase, results in 5 time values for each half. In the first half of the baseline phase (A), the median time value is 3 with a corresponding median blood pressure value of 170. For the second half of the baseline phase (A), the median time value is 5 with a median systolic blood pressure level of 170. Drawing a line connecting the two coordinates completes the procedure. It is clear from the trend lines that no consistent upward or downward trend exists in the initial baseline measures, an important consideration in interpreting the treatment effects. It is also clear through inspection of the trend lines that a systematic decrease and increase in systolic blood coincides with the presentation and removal of the intervention, respectively. An important consideration in establishing trends is to examine the variability within and between each phase. The trend ranges (calculated in the same manner as range lines [23]) shown in Fig. 4.2 suggest that the variability is decreasing during intervention, as well as when the treatment is removed. The reduction in variability, if accurately measured in this scenario, may simply reflect the adjustment of the patient to the presentation and removal of the medication. More measures would be useful in testing this notion, as well as determining the limits of the effectiveness of the medication in further reducing systolic blood pressure.

Figure 4.3 illustrates an A-B design containing three patients. In this example, all of the data were taken from patients in a study conducted at a primary care site. Note

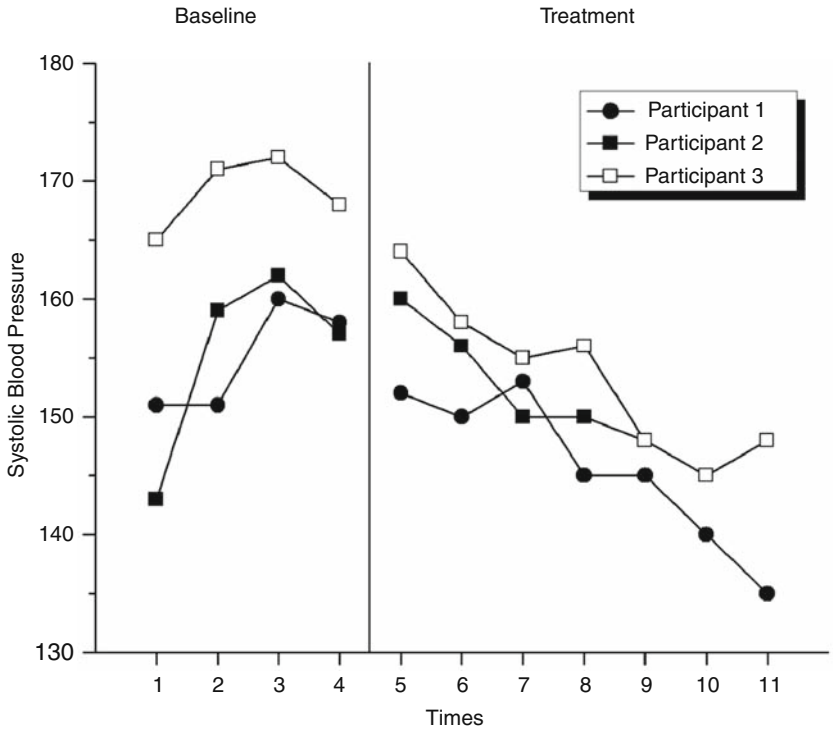

Fig. 4.3 Illustration of an A-B design, with three patients

in this case that although the variability in systolic blood pressure for each patient is modest, it would have been useful to have more baseline measures to further assure the stability of the measures, especially given the lack of a return to baseline condition. However, ethical issues must be considered when removing treatments that are beneficial for patients. Also, note the decline in systolic blood pressure across the treatment phase for the three patients, suggesting that the effectiveness of the medication is not unique to any given individual patient. If the purpose of the study was to determine the generalized effectiveness of the medication, it would have been useful to have more patients. It is also uncertain whether systolic blood pressure levels would continue to decline with additional treatments. Finally, a follow-up would have been useful. Depending on the purpose of study, the previously mentioned statistics may be applied to these data. For example, it may be useful to display a single trend line and range line (computing these values based on all three patients) for baseline and treatment, especially as the data suggest little variation among the patients.

Figure 4.4 displays data from an alternating treatments design, which was discussed in Chapter 2. In this study, the physician was interested in the effectiveness of an increased dosage of the current insulin regime on reducing Hemaglobin A1C. Of note in this study is that the researcher had data to indicate that the baseline (A) could be established or stabilized with only two measures. However, in general, it

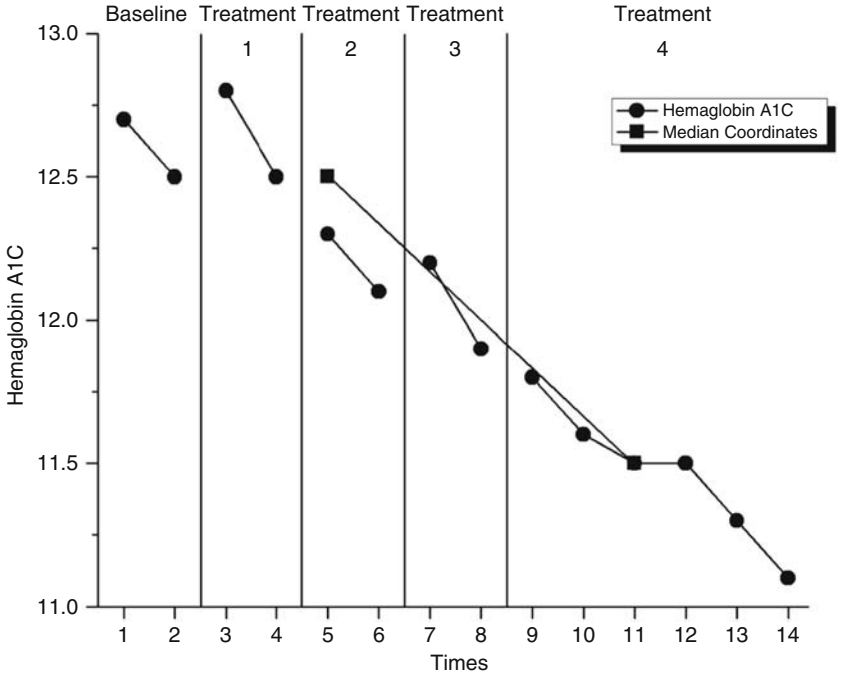

Fig. 4.4 An example of an alternating treatments design

is unlikely that a stable baseline can be established with two measures. It is recommended that there should be a minimum of three observations, and obviously more, if there is considerable variability in the measurements [19]. Referring to Fig. 4.4, notice the decline from the first to the second baseline measure. Of consideration is whether this is a trend and whether it would contaminate the intervention. Nonetheless, the major purpose of this study was to determine the impact of increasing levels of the intervention on Hemaglobin A1C, and it is clear that a trend exists in the data. Based on all of the data (i.e., the baseline and the four treatments) the split-middle method was used to create the trend lines. It appears to fit the data quite well, and perhaps range lines might be useful in this case to show variability. Finally, insulin dosage levels systematically increased across the time course of the study. Another point of consideration when creating a study is to determine whether the results could be replicated if administration of the drug dosage levels occurred randomly. A more in-depth discussion of this topic will be presented later in this Chapter.

Should visual analysis be used? Considerable discussion and debate have surrounded the use of visual analysis [2, 5, 16, 19, 21, 23, 28, 29]. Franklin, Allison, and Gorman [23] argued that one should use caution when interpreting graphs using visual inspection. One of the most crucial assumptions of visual analysis is that the observer can provide an accurate causal inference of the relationship depicted in the graph. Unfortunately, studies have shown that there is little agreement among observers of the same graphs, even if the observers were trained in the use of techniques of interpretation (See Franklin et al. [23] for a summary of some of this early research). Graphed data are vulnerable to confirmatory biases if care is not taken in the scaling of the graph [28]. In addition, confirmatory biases can occur with Type I errors, such as when observers or researchers see what they want to see, especially if the data are serially dependent, and assume an effect exists when in reality it does not [21, 29, 30]. Simply changing the scale values on the axes can make the data subject to misinterpretation, as amply demonstrated by Huff [31] in his popular text, *How to Lie with Statistics.*

Kazdin [32] and Cooper et al. [15] have argued that the use of visual inspection should be restricted to large and reliable effects because the interpretation of large effects are less susceptible to misinterpretation, and they possess more clinical and social significance than small effects. Of concern for consideration of this recommendation is the size of the effect as well as the variability in the data. If the size of the effect is large, the visual interpretation remains suspect if there is high heterogeneity. Cooper et al. [15] further argue that this approach leads to fewer Type I and more Type II errors in data with small effects. In contrast, Franklin et al. [23] point out that in the long run, these small but reliable effects may produce more permanent and important patient effects that are overlooked by visual analysis. In other words, the commission of Type II errors is not a benefit of visual inspection but rather a detriment. Unfortunately, it is not uncommon in more traditional statistically based forms of research to find very small but significant effects where the likelihood of clinical application would be miniscule. It is also important to recognize that finding statistical significance does not necessarily mean that every patient improves with

the treatment; therefore, single subject research may be necessary to determine if the treatment is successful for a given individual [33].

Kazdin [2] stated that the inconsistencies and subjectivity in decision making using visual analysis were possibly a result of the failure of researchers to establish a systematic set of rules to follow during the process. As mentioned earlier in this section, it is critical that a consistent approach be used when visual inspection is employed, especially given the recommendation that visual inspection be the primary, if not sole, method of analysis in single subject research [15, 24, 34]. In summary, Cooper et al. [15] and the aforementioned recommendations should serve as a useful guide. Although further research will be needed to resolve the visual analysis debate, if these recommendations are followed, it is likely that the reliability, validity, and accuracy of visual inspection will improve as a method of analysis. Finally, partly because of the usefulness to physicians and researchers, it is clear that visual inspection will continue to be used, regardless of its empirical status. Therefore, it is essential that the method be improved and supplemented by other means. The next and final section of this Chapter will examine the usefulness of inferential statistics to supplement visual inspection, an approach recommended by Franklin et al. [23] and others (e.g., Houle [21]).

Inferential Statistical Analysis

For the evaluation of data from a single subject design, it is important to understand some of the basic concepts of the classical statistical approach to inferential testing. First, for context, it is necessary to make a distinction between descriptive statistics and inferential statistics. Descriptive statistics refer to the quantification of the summary information from the studied sample of patients. It is common practice to summarize a sample or group of measurements by providing measures of central tendency (mean, median, and mode), along with measures of dispersion or variability (range, standard deviation, and variance). Inferential statistics refer to techniques of inferring population characteristics from the sample data. A population is defined as a set that share at least one characteristic in common, and a sample is simply a subset of the population. Of course, in order to have a representative sample, it is important that the sample be randomly selected, in that every element within the population has an equal chance of being selected. For inference or estimation to population values from the sample, these population values are termed parameters. Inferential statistical testing is typically subdivided into further classifications, with parametric testing and nonparametric (or distribution free) testing as one of the subdivisions. Parametric tests are designed to test the distributional characteristics of the population based on the sample values. In contrast, nonparametric tests are distribution free, in which no assumptions are made about the form of the sampled population. The relevance of these statistical concepts to single subject research has raised considerable discussion [28]. A number of inferential tests that have been

used with data collected from single subject designs, and all of these tests have advantages and disadvantages. At this time, randomization tests seem to hold the most promise.

Randomization Tests

Edgington [1] has strongly advocated for the use of randomization tests. Edgington [1] noted that statistical tests can be used if there is treatment randomization and random sampling is not a necessity. Randomization tests are distribution free or nonparametric. The test statistic (e.g., *t* or *F*) is calculated based upon the observed data. The significance of the test statistic is based on the number of ways (i.e., permutations) in which the data can be ordered. Finally, the test statistic is computed for each order and the probability of the treatment relative to the permutations is used to determine statistical significance. Edgington [1] has argued that for inferential testing with single subject designs, randomization tests are the sole method. Furthermore, randomization tests are appropriate when there is serial dependency in the data, as sometimes it is expected with single subject designs. Krishef [20] noted that the disadvantages included: the inability to generalize to a population; that in multiple treatment studies there may be carry-over effects; and the laborious calculations required for determining the number of permutations. The latter disadvantage is no longer a serious concern given the advent of recent technological and computing advances [21]. For a more in-depth examination of this approach see Edgington [1] and Houle [21], along with Krishef [20] for a computational example. See Bulte and Onghena [35], Onghena and Edgington [34], and Todman and Dugard [36] for the application of randomized trials to medicine.

Nonparametric Smoother

As a complement to visual inspection, Janosky [37], Janosky, Al-Shboul, and Pellitieri [38], and Janosky, Pellitieri, and Al-Shboul [39] discussed the implementation of a nonparametric smoother for use with single subject designs. The nonparametric smoother is applied to the collected series of data points, and the analysis leads to a smoothing of the function by separating an actual or true process or model, from error or noise in the collected data. The nonparametric smoother does not require the statistical assumptions of parametric testing, and it can be used as a supplement to visual inspection. Empirical tests show that the smoother works well with linear models, and it avoids some of the problems associated with visual inspection (e.g., distorted plots, broadened or narrowed axes and inappropriate use of scales). The major disadvantage is limited applicability for cyclical models, when the number of collected data points is not large. See Janosky [28, 37], Janosky, Al-Shboul, and Pellitieri [38], and Janosky et al. [39].

Celeration Line Methods

Krishef [20] described two celeration (acceleration or deceleration) methods for determining trend lines. The split-middle method has already been described and applied to the data depicted in Fig. 4.2. This method uses medians to determine trend lines, whereas the second method, called celeration line, uses means to plot the trend lines. For both methods, based on the binominal distribution, the purpose is to determine from the baseline data whether the treatment data can be predicted, or if the rates of change differ? Both methods require a minimum of 10 observations for the baseline and a minimum of 5 for the treatment phase. The major advantage of the celeration line method is that it provides an estimate of the trends, if any are in the data. As with the nonparametric smoother, celeration methods may be more useful as a descriptive adjunct or aid to visual inspection. The disadvantages include limited applicability if the data are auto-correlated (the binominal test requires that the observations be independent), difficulty in interpretation when trends lines are approaching asymptote during the baseline, and the meeting of the minimum requirements for baseline and treatment measurements may not be practical with some patients. See Cooper et al. [15], Franklin et al. [23], Houle [21], Janosky and Al-Shboul [40], and Kazdin [2] for a more in-depth discussion, as well as computational examples.

Sheward's Two Standard Deviation Band

If the celebration line method cannot be used due to practical concerns, a possible alternative is to use Sheward's chart procedure [20]. The significance test is based on determining whether two successive observations fall outside the band of plus or minus-two standard deviations. The advantages of use include straight-forward computations and general application to any single subject design. The disadvantages are many, including the necessary assumption of random variation, no autocorrelation in the data, stable baselines, and no trends in the data.

Bartlett's Test

Bartlett's test allows for a determination of whether an autocorrelation exists in the data. The computation of the correlation is based on lagged values (i.e., a serial correlation) and can be used when data are collected in a sequential manner. Examples are available through the works of Krishef [20], Kazdin [2], McGuigan [7], Pittenger [41], and Kirk [42].

Mann-Whitney U

The Mann-Whitney U test is a nonparametric test that can be used for analyzing single subject research, in which each subject receives two or more treatments or

interventions. Statistically significant differences between the treatment conditions can be analyzed. The test also requires treatment randomization. The advantages of the Mann-Whitney U include limited statistical assumptions and ease in computation and interpretation. With the presence of treatment randomization, serial dependency is not an issue of concern. The disadvantages of the Mann-Whitney U include the lack of appropriateness for designs, where treatments are irreversible and treatment carry over effects are suspect [20]. However, if more than one patient is used in the study, it may be possible to control and analyze for carry over effects by counterbalancing the order of the treatments. See Krishef [20] and Kirk [42] for more detail and computational examples.

Revusky's R_n Statistic

Randomization tests assume independence of observations. If treatment effects are irreversible and it is not possible to remove the intervention and return to baseline, the researcher may decide to use the A-B or multiple baseline design as an alternative. Revusky [43] developed a statistic (R_n) that can be used to analyze data generated from these designs. A minimum of four baseline measures are required before using the statistic, and the intervention must be randomly assigned and given only once. The statistic can be used with all of the variations of the A-B designs (i.e., across subjects, across behaviors, and across situations). This test evaluates the statistical significance between the treatment and untreated phases. The strengths of this test include the applicability when treatment(s) cannot be withdrawn, ease in calculation, and the superior level of sensitivity to detect effects, as compared to the Mann-Whitney U [20]. The major weakness is the necessity for the intervention randomization requirement, since this is sometimes difficult to meet due to practical obstacles (e.g., a particular patient requires immediate treatment, thus failing to meet the requirement of randomization of treatments to patients). The other possible weakness is the inability of the researcher to obtain the minimum requirement of four baseline measures. For additional information, please see Kazdin [2], Houle [21], Krishef [20], and Revusky [43].

The W Statistic

The W statistic has been discussed in detail by Krishef [20]. Similar to the application of the R statistic, the W statistic is appropriate to use with multiple baseline designs. In contrast to the R statistic requirements, the W statistic does not require that the treatment be ended after each intervention. Randomization is necessary for determining the order in which the patients (also applicable to across behaviors and situations) receive a treatment. For the W statistic, a comparison is made between the baseline and treatment for each individual patient. The number of permutations (based on the number of patients, behaviors, or situations) drives the W statistic, and statistical significance is then determined. The W statistic is

essentially a randomization test. As the baseline and treatment phases for each patient are compared, the advantage of W statistic is more applicable in an applied setting focusing on the effectiveness of the treatment. This approach also does not require the immediate termination of the treatment as the R statistic requires. The disadvantages are similar to any randomization test (discussed under the randomization tests). For a computational example see Krishef [20].

The C Statistic

According to Krishef [20], the C statistic can be used in determining whether there are abrupt changes in level, but only when there are minimal changes in slope or direction. The C statistic can be used to test the stability of the baseline, as well as comparing the baseline with the treatment phases. The latter is accomplished by determining whether the slopes are different for the baseline and treatment phases. This statistic requires a minimum of 8 observations. The advantages of this statistic are that it can be used to determine the effectiveness of the treatment with 8 or more observations, even though the data may be serially dependent. Furthermore, the statistic is simple to calculate especially relative to the more complicated and consuming analyses dealing with time series data {e.g., ARIMA (auto-regressive integrated moving averages; Houle [21])}. One disadvantage includes the failure of the statistic to detect abrupt changes in direction of the function. A second disadvantage is the effect on statistical power. Simply having more data points when the baseline and treatment are combined for analysis may lead to statistical significance, whereas only analyzing the baseline may not. See Jones [44] and Krishef [20] for additional discussion.

What should the role of inferential statistics be in single subject design research? There is considerable diversity of opinion regarding the utility of inferential statistics in single subject research. Some have relied largely on visual analysis [15], arguing that clinical significance requires large effects that can be easily interpreted using visual analysis, and statistical analysis may be misleading if small effects are found to be significant [2, 4]. Barlow et al. [4] further state that one may find statistical significance with considerable error, which may indicate the treatment is effective for some individuals and not others. Essentially, trends and intra-subject averaging may mask the variability in the data. Finally, Kazdin [2] has argued that because of the pervasiveness of statistical inferential testing in the sciences, researchers may fail to conduct single subject research on a promising topic or change the design because there is no statistical analysis available to evaluate the data. Furthermore, Kazdin [2] has discussed the debate regarding whether inferential statistics should be used and whether the data from single subject designs meet the assumptions of parametric statistics. Kazdin [2] states that statistics can be used when baselines are unstable, whether the intervention is reliably different from the baseline, when there is considerable intra-subject variability, and during the investigation of new areas where weak effects may be detected, but show some promise for future research. Kazdin

[2, 14] has recommended the use of parametric statistics under these conditions, if the assumptions of parametric statistics can be satisfied. Unfortunately, it is rare that these assumptions can be met because of the inherent characteristics of single subject research. A more conservative approach is to use statistics as a supplement to visual analysis [16, 20, 21] and possibly to restrict their use to descriptive statistics, as the requirements for descriptive statistics are more readily met for single subject designs [45]. The argument is that statistics can be used to confirm what is presented in a graph [1, 20]. Unfortunately, even using this conservative approach can lead to invalid inferences. If parametric statistics cannot be used in single subject research, some medical researchers have suggested that data from single subject designs only be used in the early stages of development. Specifically, hypotheses can be formulated and tested later using other research paradigms [46, 47]. A more favorable approach would be to use nonparametric statistics that do not require the assumptions of parametric statistics.

Summary

The proper conduct of single subject research is essential for the welfare of patients. Sound single subject research requires the researcher to infer that the dependent variable (medical outcome) is due to the influence of the independent variable (intervention), and not to other sources (extraneous variables). Although it is unlikely that extraneous variables can be entirely eliminated from studies, it is feasible to conduct research where the influence of these variables is minimized. As a result, more confidence can be placed in the causal relationship between the treatment or intervention and the outcome. In order to minimize the role of extraneous variables, it is important to rigorously apply the techniques of control (i.e., elimination, constancy, balancing, counter-balancing and randomization) in the design of the study. Furthermore, it is important to establish the integrity or fidelity of the independent variable, as well as its reliability, validity, and accuracy. Reliability, validity, and accuracy must also be established for the dependent variable, with particular attention paid to the benefits that the patient may receive from the intervention.

Although response guided experimentation is a common approach in single subject research, controversy has evolved over its use, largely because of the role of the physician or researcher in influencing the outcome of the study. Referencing strengths, the use of response guided experimentation may bestow benefits to the patient that otherwise would not be. Response guided experimentation is a useful methodological tool and therefore, every attempt should be made to minimize the role of the researcher in the study outcomes.

In addition to assessing the quality of the data that are generated from single subject research, it is also important to consider the way in which the data are displayed and interpreted. It is common practice to graph and interpret the data using visual analysis. There are many benefits to graphing the data but research has shown that interpreting the data using visual analysis alone may be subject to human error. In

order to minimize the role of human error, it is important that proper and standardized methods be used in constructing and interpreting graphs because their use is likely to continue. Visual analysis can also be supplemented with statistical analysis, but in many cases the requirements of parametric testing cannot be satisfied, leading to the possible usage of nonparametric techniques in single subject research.

References

1. Edgington ES. Statistics and single case analysis. *Progress in Behavior Modification.* 1984; 16: 83–119.
2. Kazdin AE. *Single-case research designs: Methods for clinical and applied settings*. New York: Oxford, 1982.
3. Bernard, C. *An introduction to the study of experimental medicine.* New York: Dover, 1957.
4. Barlow D, Nock M, Hersen M. *Single-case experimental designs: Strategies for studying behavior for change.* New York: Pearson, 2009.
5. Shadish WR, Cook TD, Campbell DT. *Experimental and quasi-experimental designs for generalized causal inference.* Boston, MA: Houghton Mifflin, 2001.
6. Ittenbach RF, Lawhead WF. Historical and philosophical foundations of single-case research. In RD Franklin, DB Allison, BS Gorman (Eds.). *Design and analysis of single-case research* (pp. 13–39). Mahwah, NJ: Lawrence Erlbaum, 1997.
7. McGuigan FJ. *Experimental psychology: Methods of research*. Englewood Cliffs, NJ: Prentice Hall, 1997.
8. Gresham FM. Treatment integrity in single-subject research. In RD Franklin, DB Allison, BS Gorman (Eds.). *Design and analysis of single-case research* (pp. 93–117). Mahwah, NJ: Lawrence Erlbaum, 1997.
9. Kazdin AE. Comparative outcome studies of psychotherapy: Methodological issues and strategies. *Journal of Consulting and Clinical Psychology.* 1986; 54: 95–105.
10. Cone J. Psychometric considerations. In M Hersen, A Bellack (Eds.). *Behavioral assessment: A practical handbook* (pp. 38–70). New York: Pergamon, 1981.
11. Johnston J, Pennypacker J. *Strategies and tactics of human behavioral research.* Hillsdale, NJ: Lawrence Erlbaum, 1980.
12. Brennen R, Kane, M. An index of dependability for mastery tests. *Journal of Educational Measurement*. 1977; 14: 277–289.
13. Suen HK. *Principles of test theories*. Hillsdale, NJ: Lawrence Erlbaum, 1990.
14. Kazdin AE. Statistical analyses for single-case experimental designs. In DH Barlow, M. Hersen. *Single case experimental designs: Strategies for studying behavior change* (pp. 285–324). Boston, MA: Allyn and Bacon, 1984.
15. Cooper JO, Heron, TE, Heward, WL. *Applied behavior analysis.* Upper Saddle River, NJ: Pearson, 2007.
16. Richards SB, Taylor RL, Ramasamy R, Richards, RY. *Single subject research: Applications in educational and clinical settings.* Belmont, CA: Wadsworth, 1999.
17. Primavera LH, Allison DB, Alfonso VC. Measurement of dependent variables. In RD Franklin, DB Allison, BS Gorman. (Eds.). *Design and analysis of single-case research.* Mahwah, NJ Lawrence Erlbaum, 1997.
18. Edgington ES. Response-guided experimentation. *Contemporary Psychology.* 1983; 28: 64–65.
19. Barlow DH, Hersen M. *Single case experimental designs: Strategies for studying behavior change*. Boston, MA: Allyn and Bacon, 1984.
20. Krishef CH. *Fundamental approaches to single subject design and analysis.* Malabar, FL: Krieger Publishing Company, 1991.

21. Houle TT. Statistical analyses for single-case experimental designs. In D Barlow, M Nock, M Hersen. *Single-case experimental designs: Strategies for studying behavior for change* (pp. 271–305). New York: Pearson, 2009.
22. Gigerenzer G, Gaissmaier W, Kurz-Milcke E., Schwartz LM, Woloshin S. Helping doctors and patients make sense of health statistics. *Psychological Science in the Public Interest.* 2008: 8: 53–96.
23. Franklin RD, Allison DB, Gorman BS. *Design and analysis of single-case research.* Mahwah, NJ: Lawrence Erlbaum, 1997.
24. Parsonson BS, Baer DM. The analysis and presentation of graphic data. In TR Kratochwill (Ed.). *Single subject research: Strategies for evaluating change* (pp. 101–165). New York: Academic Press, 1978.
25. Carr JE, Burkholder EO. Creating single subject design graphs with Microsoft excel. *Journal of Applied Behavior Analysis.* 1998; 31: 245–251.
26. Komaki JI, Coombs T, Redding Jr. TP, Schepman S. A rich and rigorous examination of applied behavior analysis research in the world of work. In CL Cooper, IT Robertson (Eds.). *International review of industrial and organization psychology.* Sussex: John Wiley, 2000.
27. Kazdin AE. Statistical analysis for single-case experimental designs. In M Hersen, DH Barlow (Eds.). *Single-case experimental designs: Strategies for studying behavior change* New York: Pergamon, 1976.
28. Janosky JE. Use of the single subject design for practice based primary care research. *Post Graduate Medicine Journal*, 2005; 81: 549–551.
29. Matyas TA, Greenwood KM. The effect of serial dependence on visual judgment in single-case charts: An addendum. *Occupational Therapy Journal*. 1990; 10: 208–220.
30. Matyas TA, Greenwood KM. Visual analysis of single-case time-series: Effects of variability, serial dependence and magnitude of intervention effects. *Journal of Applied Behavior Analysis.* 1990; 23: 341–351.
31. Huff D. *How to lie with statistics*. New York: W.W. Norton, 1993.
32. Kazdin AE. *Research design in clinical psychology*. Boston, MA: Allyn & Bacon, 1992.
33. Johnson CM. Validating case studies in family medicine: Single-subject research designs. *Family Practice Research Journal.* 1984; 4: 27–35.
34. Onghena P. Edgington PS. Customization of pain treatments: Single-case design and analysis. *Clinical Journal of Pain.* 2005; 21: 56–68.
35. Bulté I, Onghena P. An R package for single-case randomization tests. *Behavior Research Methods.* 2008; 40: 467–478.
36. Todman J, Dugard P. *Single-case and small n experimental designs: A practical guide to randomization tests*. Mahwah, NJ: Lawrence Erlbaum, 2001.
37. Janosky JE. Use of the nonparametric smoother for examination of data from a single-subject design. *Behavior Modification.* 1992; 16: 387–399.
38. Janosky JE, Al-Shboul QM, Pellitieri TR. Validation of the use of a nonparametric smoother for the examination of data from a single-subject design. *Behavior Modification.* 1995;19: 307–24.
39. Janosky JE, Pellitieri TR, Al-Shboul QM. The need for a revised lower limit for the nonparametric smoother. *Statistics and Probability Letters*. 1997; 32: 269–72.
40. Janosky, JE, Al-shboul, Q. Statistical analysis of single-subject designs. Physical Therapy. 1995; 75: 157–8.
41. Pittenger, DJ. *Behavioral research: Design and analysis*. New York: McGraw-Hill, 2003.
42. Kirk, RE. *Statistics: An introduction.* Fort Worth, TX: Holt, Rinehart and Winston, 1990.
43. Revusky, SH. Some statistical treatments compatible with individual organism methodology. *Journal of Experimental Analysis of Behavior.* 1967; 10: 319–330.
44. Jones, PW. Single-case time series with Bayesian analysis: A practitioner's guide. *Measurement and Evaluationin Counseling and Development.* 2008; 36: 28–39.
45. Kratochwill TR. (Ed.). *Single subject research: Strategies for evaluating change*. New York: Academic press, 1978.

46. Furedy JJ. Commentary: On the limited role of the “single-subject” design in psychology: Hypothesis generating but not testing. *Journal of Behavior Therapy and Experimental Psychiatry.* 1999; 30: 21–22.
47. Reboussin DM, Morgan TM. Statistical considerations in the use and analysis of single-subject designs. *Medicine and Science in Sports and Exercise*. 1996; 28: 639–644.
48. Silvestri SM. How to make a graph using Microsoft Excel. Unpublished manuscript. Columbus, OH: The Ohio State University, 2005.