Chapter 3
Methodological Framework for Single Subject Designs

This chapter presents a methodological framework for single subject designs. In particular, the historical roots of research methodology are examined, including a discussion as to possible barriers to application that resulted in the underutilization of single subject designs. Included is a comparison of the strengths and challenges in the context of internal and external validity. Compared to traditional between-group designs, single subject designs have comparable or stronger internal validity but are more limited in some aspects of external validity; that is, the single subject design may provide more definitive conclusions, but it can be more difficult to generalize those conclusions to other participants or patients. Strategies for overcoming these limitations are examined.

Historical Roots

Although clinical practice focuses on the individual, biomedical research has primarily focused on the study of groups, including the evaluation of biomedical interventions implemented with groups of patients. The considered gold standard within biomedical research, the randomized controlled trial (RCT), is most often used to evaluate interventions for groups or cohorts of patients or subjects. Even though the RCT, considered as an experimental design, has typically taken precedence over the other research methodologies, including the single subject design, all methodologies have inherent strengths and weaknesses. For biomedical researchers, the best course for increasing scientific understanding of relevant phenomena revolves around the utilization of a variety of methodological designs, with the research question of interest determining the choice of the design.

This section provides an examination of the historical roots of the single subject design to highlight the importance of use, while also clarifying why it has been underutilized in biomedicine. Currently, single subject designs are being employed more frequently and provide a number of opportunities for improving direct patient care, as well as answering important biomedical research questions [1].

J.E. Janosky et al., *Single Subject Designs in Biomedicine*,
DOI 10.1007/978-90-481-2444-2_3,

Individual-Focused Designs

Whereas between-group designs became more utilized after several statistical discoveries in the 1930 s, informal single subject design research began to propagate nearly one hundred years prior, in the 1830 s. Most early research involving single subjects was conducted within the budding field of neurophysiology. In particular, Hall and Flourens began conducting experimental ablation studies, which examined the physiological and behavioral effects of destroying or removing various brain regions [2]. Capitalizing on their earlier research, Broca described the relationship between language deficits and localized brain lesions observed through post-mortem examinations [3].

In the research area of sensation and perception, the single subject design was frequently employed; for example, Fechner examined the minimum thresholds necessary for perception [4]. This work by Fechner on just noticeable differences (JND) was unique in the use of statistics to quantify the minimum necessary increase in stimulus intensity needed for discernment. Later experiments by Ebbinghaus, examining memory, and Pavlov, examining classical conditioning, or associative learning, were similar in design – relying extensively on multiple observations of single subjects [5, 6].

Although several examples of rigorous single subject experimental design studies have been noted, the early study of single cases was relatively informal, particularly in the applied setting. Case studies are detailed accounts of single cases, and they differ from single subject design studies in that the investigator typically exercises less control and may not rigorously collect and analyze quantitative data. During the late 1800 s and early 1900 s, case studies were the primary method of clinical investigation. For example, neurologist Jean-Martin Charcot's early case reports helped to document conditions such as Charcot-Marie-Tooth disease, multiple sclerosis, and Parkinson's disease [7]. Charcot became primarily interested in studying patients suffering from "hysteria" or physical symptoms with no neurological basis (commonly referred to as somatization disorders or conversion disorders today). Charcot mentored a number of notable psychologists, including Sigmund Freud, the quintessential case study investigator. Freud's evolving theories of psychopathology drew heavily on case material obtained from his patients, and he published several lengthy case reports. Although Freud may have been most notable, this methodology was characteristic of most clinical psychologists in the early 1900 s. Of course, case studies suffered from a number of major limitations, in that they rarely relied on data, systematic observation, or experimental control. Those using case studies often made bold claims of treatment effectiveness or postulated a number of unsupported inferences in their theories. Inevitably, researchers became disenchanted with case studies. Perhaps because case studies were much more common than rigorous single subject design studies, researchers tended to disregard individual-focused investigations altogether, shifting increasingly toward group-level designs. Thus, it may be argued that the paradigmatic shift away from individual-focused research could be typified excising the weaknesses of the case study at the expense of important single subject design research. This paradigmatic shift was also facilitated by statistical advances most easily applicable to between-group designs.

Group Experimental Designs

Eventually, scientists became increasingly interested in studying human (as well as interspecies) variation [8]. Researchers began to note that many important human attributes, skills, and abilities varied along a standard normal or "bell" curve, and the need for selecting qualified military recruits in the 1900 s led to increased focus on intelligence testing [9]. The researcher's locus of observation had shifted from intra-individual to inter-individual differences.

This changing focus in methodology was also catalyzed by several important statistical discoveries. Pearson and Galton worked to advance the field of descriptive statistics, through their work on correlation, regression, and chi-square tests [10]. Ultimately, these techniques were expanded, with correlational techniques providing the foundation for later work on factor analysis, which was used predominantly in studies analyzing individual differences in personality traits and cognitive abilities. Thus, the development of descriptive statistics aided the quantification of individual differences.

During the early 1900 s, the initial publications on inferential statistics also began to appear. While working for the Guinness brewing company, Gosset began developing formulas for monitoring quality assurance of brews, and drawing heavily on the correlational work of Pearson in discovering formulas for comparing group means [11]. Although his statistical work was considered a part of trade secrets of his employer, in 1908 Gosset detailed his findings on t-tests (publishing under the pseudonym "Student" to protect himself from legal liability). These t-tests allowed for comparing a sample mean to a population mean or to other samples. Yet, the importance of t-tests was not fully realized until the later work of R. A. Fisher. In laying the foundation for inferential statistics, Fisher documented how probability could be used to determine the reliability or significance of results [12]. In particular, for t-tests and other related statistics, probability values could be ascertained describing the odds that observed mean differences could be obtained by sampling error, the chance variation that occurs across samples. Researchers now had a method for determining whether groups differed based on the probability that mean differences were due to sampling error and this statistical advancement may have led to greater reliance on the between-group methodology. The statistical power of a study, or its ability to detect an effect when it is present, increases with sample size; that is, larger N studies are better able to detect differences yielding more accurate results. The findings of studies with small sample sizes were increasingly criticized, as a result of this advancement. Publishing trends in the 1930 s documented a rapid shift away from small-sample studies toward large-sample studies, drawing upon inferential statistics [13, 14]. Too often replacing the approach of controlling for variation through precise experimental control, researchers began averaging individual differences through increasing sample sizes and statistical techniques.

Return to Single Subject Designs

A number of researchers hold that single subject designs can overcome some limitations inherent in between-group designs [15]. Ethically, between-group designs

were disadvantaged when using control or waitlist conditions that denied some patients useful treatments. Because the between-group design relies on large samples to average out (i.e., sum over) individual differences, several pragmatic concerns also arose. Specifically, at times it is difficult to find a large number of patients who have unique demographics or suffer from rare diseases. Furthermore, large N studies can be time consuming. One of the consequences of the time consuming nature of large N research is the difficulty in studying public health crises, for example. Additionally, the exorbitant financial costs of large-sample research often limit who is able to conduct such projects, at times risking an ethical dilemma with the linking of the researcher and the funder in mutual vested interests in the results. For example, funding from pharmaceutical companies is often needed to conduct the multi-million dollar research necessary for evaluating the same drugs those companies produce [16].

Beyond the ethical and pragmatic limitations of between-group designs, there are also methodological reasons for using the single subject approach. Basically, the two approaches have different methods for handling variability in outcomes. For the between-group design, rather than attempting to bring differences in outcome under experimental control, the researcher averages out differences in outcomes by using large samples. Within this methodology, the researcher seeks statistical control over error, rather than experimental control to reduce error. This strategy is problematic for two reasons: (1) statistical power and sample size are related, with larger samples at times leading to significant but very small effects with little pragmatic value and (2) it discourages the researcher from strategically modifying treatment (i.e., response guided experimentation) that may positively impact most if not all the patients. In a between-group design, a treatment condition can produce a statistically significant outcome that is more advantageous than a control condition, but this difference is based on mean differences, that is, the treatment could benefit some patients but not others.

In contrast, the single subject design methodology permits the researcher to exercise extra control over the intervention. If a participant does not respond to a particular treatment, a desired effect may be achievable through a modification or change in the treatment through response guided experimentation (See Chapter 5 for further discussion of this approach).

Beginning in the 1930 s and expanding rapidly in the 1950 s and 1960 s, Skinner helped pioneer small-sample research. Given the above criticisms of between-group studies, Skinner emphasized studying the individual to determine lawful models of behavior. He drew heavily upon animal research, often using pigeons or rats, to uncover fundamental learning principles that could then be applied to humans [17–19]. Inevitably, similar procedures for modifying behavior were applied to individual human subjects. Within the realm of applied behavior analysis, single subject design studies began examining methods for modifying behavior of individuals with diverse psychological problems, including stuttering, learning disabilities, mental retardation, and psychotic symptoms [20].

More recently the single subject design methodology has extended beyond the fields of psychology and education to biomedicine; for example, single subject

designs may be nested within larger clinical trials to increase compliance and answer more detailed questions [21]. Single subject designs are particularly useful for answering questions regarding rare diseases, side effects, unique populations, emergency situations, and isolated environments, in which between-group designs would be unfeasible or impractical [22, 23]. This methodology is also particularly suited for primary care practice-based research, where practitioners can tailor individualized treatments to improve outcomes [23, 24].

Sources of Internal Validity Threats

Internal validity refers to the strength of inferences that can be made regarding the relationship between two variables. Depending upon the methodology employed, at times the inferences may be causal. Within the context of biomedical research, internal validity typically refers to the extent to which observed outcomes can be attributed to the intervention. For example, consider a psychiatric pharmaceutical trial for treating major depression. If the methodology of the study supports strong conclusions about the ability of the treatment medication to lessen depressive symptoms, then it may be concluded that the study has internal validity. Internal validity is weakened to the extent that the results can be challenged by methodological pitfalls or alternative explanations. For example, if the study did not include proper controls, the causal effect of the specific medication on the outcome could be questioned. Basically, the internal validity of any research finding, including biomedical findings, can questioned because of the inherent methodological limits of the research design being used. Therefore, it is best to view internal validity on a continuum, with each methodological approach containing strengths and weaknesses.

Causation

In order to assess the internal validity of a study, it is foremost to understand what is meant by "causation". Hume was the first to articulate a precise definition of causation, noting that a causal relationship could only be inferred when three conditions were present: temporal precedence, covariation, and no plausible alternatives [25].

Most importantly, the causal variable must precede the effect (i.e., temporal precedence). In a drug trial, for example, the observed effect is noted to only occur after the treatment has begun. Typically, establishing temporal precedence in experimental studies, such as single subject and between-group designs, is relatively straightforward, assuming the experimental manipulation occurs before the change in symptoms. In contrast, causality is more difficult to establish in non-experimental research (e.g., quasi-experimental and systematic observation studies) because it is difficult to establish temporal precedence.

Secondly, for a determination of causality, there must be covariation between the cause and effect; that is, the effect must be more likely to occur when the presumed causal variable is present than when it is absent. For example, medication use covaries with a reduction in depressive symptoms if symptoms decrease more when medication is administered than when it is not administered. The magnitude

of covariation is indicated using various measures of effect size, such as Pearson's *r*, Cohen's *d*, or other statistics [26]. However, often more concrete examples, such as changes in actual recorded values or well-constructed graphics, may be just as informative. Finally, causation can only be inferred if there are no credible alternative explanations. For example, if a psychiatric drug and a placebo similarly impacted depressive symptoms, it could not logically be argued that the drug had any specific antidepressant effects. Generally, of the three criteria, ruling out alternative explanations is the most difficult to meet.

Within the context of medical research, Hill introduced a list of nine points researchers should consider in evaluating evidence for causation, including the strength of the relationship, consistency across contexts, specificity of effects upon unique outcomes, temporal order, biological gradient or dose-response relationship, theoretical and biological plausibility, coherence with historical evidence, supplemental experimental evidence, and analogous findings for related interventions [27]. Other researchers have proposed similar lists, and researchers frequently choose a subset of the nine points as criteria for evaluating causal assertions in research studies [28–30].

Properly designed and executed, single subject designs can be useful in providing evidence for internal validity and may be particularly useful within primary care practice-based research [23, 24]. Specifically, experimental control may allow for the determination of large effects. Consistency across situations can be determined by using multiple baselines. Changing criteria designs can be implemented to assess the specificity of interventions upon particular outcomes. Multiple phases, involving the titration of dosages, can also be used to demonstrate a dose-response relationship. Thus, because the single subject design is often more dynamic, flexible, and customized than the between-subject design, the single subject design may be able to provide more credible evidence of internal validity than the between-subject design. However, in order for single subject researchers to establish internal validity, it is important that potential threats to internal validity be recognized and controlled when planning their research studies.

Sources of Threats to Internal Validity

This section includes a primer on the well-recognized threats to the internal validity of research studies in general [15, 31–33]. In subsequent sections, more information will be provided on how these threats are likely to occur in the between-groups and single subject designs (Table 3.1).

Mortality

Mortality threats refer to a collection of concerns surrounding patient screening, death, or drop out. In clinical trials, researchers frequently screen patients prior to selecting them for the study, with examples including length of time since diagnosis, severity of symptoms, comorbidities, or demographic features. Although selection

Table 3.1 Threats to internal validity

Threat	Description
Mortality	The inflation of an observed effect due to participant drop out, non-random selection, or the omission of select trials.
Regression toward the mean	No measure is perfectly reliable, so extreme scores generally normalize over time, generating spurious effects.
Maturation	An observed change is due to developmental changes rather than the experimental intervention.
History	An observed effect is due to a historical event rather than the treatment.
Testing effects	Rather than a controlled intervention causing changes, the measurement procedures themselves unintentionally alter future scores.
Instrumentation	Unintended changes to the measurement instruments may impact changes in the outcome measures.
Withdrawal reactions	When interventions that produce tolerance are withdrawn, they may produce side effects that mimic or aggravate the original condition, exaggerating the appearance of treatment effects.
Social-cognitive effects	Social interactions with investigators or other participants can foster changes in thinking or behavior that impact treatment effects.
Residual confounding	Because measurement instruments contain error, any effort to statistically or methodologically control for internal validity threats and other confounds will be imperfect.

criteria invariably impact the external validity, or generalizability, of results, they may also impact the internal validity of results when screening procedures are used to select patients who have an elevated probability of biased responding to the treatment. A clear example of this was shown in an SSRI study by Dimidjian, Hollon, and Dobson [34] in which patients were excluded from the study if they had failed to respond favorably to a trial of paroxetine within the past year. This most likely biased the results by only including patients with a greater probability of responding favorably [34]. Screening effects can occur in between-group and single subject designs, although screening may be more likely in large experimental designs, such as randomized clinical trials.

Among patients selected for the study, some may drop out or, unfortunately, die. Drop out, particularly noninformative drop out, can pose substantial limitations for the internal validity of clinical trials. If drop out rates vary across experimental conditions or occur for different reasons, namely informative drop outs, observed treatment effects may be due to individual differences between patients, rather than to the experimental manipulation. For example, in a medication trial, patients in the treatment group may be more likely to drop out than those in the placebo group, due to an increased level of side effects. Patients opting to continue with the experimental medication may be above average in terms of level of responding, making it difficult to compare them to the control group. For single subject designs, drop out and death are probably less likely to occur. Furthermore, because the single subject

design incorporates the possibility of changing the treatment during the experiment, the researcher can more easily respond to adverse events, such as side effects, by quickly modifying treatment. For example, in the context of primary care, this might involve altering a dosage or prescribing a secondary medication to manage a side effect. Occasionally single subject design studies have been nested within larger clinical trials, and they have been shown to dramatically reduce drop out [21].

Regression Toward the Mean

When measures are administered across two or more time points, initial scores that are extreme tend to regress toward the mean. In essence, high scores are likely to decrease and low scores are likely to increase. This statistical reality can create the appearance of treatment effects, when in fact there are none.

All scores represent the sum of two components, true variance and error variance. For example, any patient's fasting blood glucose level would be caused by their stable level of glucose as well as erroneous factors, such as measurement error (i.e., accuracy of the glucometer) or day-to-day variation (e.g., postprandial versus preprandial measurements). One possible reason for extreme scores is error variance; that is, extreme scores are due in part to uncontrolled, unmeasured, or "chance" variation. Because this variation is not systematic, it is likely to lead to reduced scores on a later re-test. Regression toward the mean, therefore, is a problem for studies examining change over time, when patients have been screened to score high on some diagnostic measure, such as having elevated glucose levels. Any symptomatic reduction could be due in part to regression rather than treatment. In a randomized experimental design, the inclusion of a control group aids in minimizing this threat; however, the problem is that regression may be disparate between the experimental and control groups. If the treatment group has greater initial symptom severity than the control group, patients may be more likely to drop out of the control group, and the apparent treatment effects will be inflated.

Regression can also lead to limitations for single subject designs. Regression may create difficulties for establishing a stable baseline prior to treatment. For example, a patient's level of depression may continue to gradually decline before treatment is introduced. This problem can be overcome by increasing the baseline period, though this option may not be practical. An additional problem arises for the simple A-B design, where symptom reduction during phase B may be due to regression. This threat is less noteworthy when symptom reduction occurs steeply at the introduction of treatment. Furthermore, regression can be overcome by using a reversal design, in which treatment is withdrawn and then re-implemented when feasible. In fact, because the single subject design can include several reversals and is designed to increase control, this methodology can provide significant advantages for countering the threat of regression. In the case of randomized clinical trials, repeated reversals may be expensive and impractical, so single subject trials offer a pragmatic alternative for addressing regression threats.

Maturation

An observed effect within a study could potentially be explained by naturally-occurring developmental processes within the organism. The most general type of maturational threat involves aging itself, though specific developmental changes in perceptual skills, cognitive abilities, social skills, emotional functioning, strength, and metabolism are worth considering. These threats are particularly important for long-term studies or studies involving groups undergoing rapid developmental changes, such as children, older adults, pregnant women, and people with degenerative diseases. For between-group experimental designs, this threat is important to consider when groups differ on major demographic variables, such as age, sex, gender, ethnicity, race, or socioeconomic status, which are intertwined with developmental variables. In biomedical studies, more specific variables need to be considered, such as initial group differences in the severity or likely course of the illness (e.g., allele frequency, ethnic differences, duration of disease, etc.).

Thus, it is important for researchers to measure these variables and attempt to ensure that patients are equally matched across groups. Unfortunately, the number of potential extraneous variables can be quite large, and whether using random- or matched- assignment, it can be difficult to ensure that patients are similar on these variables across groups. For example, consider a study comparing medication to placebo in treatment of depressive symptoms: Patients may differ on a number of health-related maturation variables that could affect responsiveness to treatment, such as the diagnostic classification (e.g. major depression versus dysthymia or Type I versus Type II Diabetes Mellitus), predominant symptoms (e.g., low mood versus anhedonic), and psychosocial underpinnings (e.g., introjective versus anaclitic), in addition to core demographic variables.

Maturational threats are important to consider in simple single subject designs (e.g. A-B or A-B-A) in which phase changes might inadvertently correspond with maturational changes. However, as the design becomes more complex or contains an increased number of reversals (e.g., A-B- A-B-A-B), the possibility that a maturational process would repeatedly correspond with the treatment effect is diminished. It is a perplexing oversight that more research has not been conducted in this regard, particularly for the study of rare medication side effects. During the past decade there has been a heated debate over whether SSRIs increase violent behavior or suicidality in some patients [16]. This question is difficult to answer using randomized controlled trials because the side effect is relatively rare, there are ethical issues surrounding the investigation of the research question, and studies with adequate statistical power would be prohibitively expensive to conduct. Dozens of case reports have been described, but maturational threats limit the internal validity of these anecdotal findings; that is, it can be difficult to determine whether increased suicidality is due to the medication or merely the progression of the depression. However, a single subject design study could be used to address this important question. For example, a physician or a practitioner could monitor increases or decreases in suicidality in response to changing doses (e.g., A-B_1-B_2-B_3), changing medications (e.g., A-B-C-D), or the addition of a secondary prescription, such

as a benzodiazepine (e.g., A-B-B-C). Although such medication changes are often conducted by physicians or practitioners as a part of treatment, they frequently lack precise measurement of symptoms or control of treatment duration. Where potential side effects may mimic the developmental course of a disorder, single subject designs afford unique opportunities for documenting and minimizing side effects. Because rare side effects are overshadowed in large, randomized controlled trials, single subject studies can have important legal and public safety ramifications.

History

The history threat refers to any event occurring at or before the time of the experiment that might confound the results. History threats are similar to maturational threats, except that the locus of the potential confounding factor is described as external to the patient, rather than as an internal developmental process. Examples include important life events, such as the death of a loved one, a marriage or divorce, changes in employment, diagnosis of a chronic disease, or an illness. Within the context of biomedical research, it would be important to examine historical variables such as, personal history of other medical problems, family health history, and presence of environmental stressors. Similar to maturational threats, history threats are important to consider in between-group studies, particularly in quasi-experimental research, where groups may differ on important historical variables. Again, the researcher should make efforts to measure and control for these historical variables, such that the confounding is eliminated or minimized. As with the benefits of controlling for maturational effects using single subject designs the same benefits apply to history effects, especially when repeated reversals are used.

Testing Effects

As the founder of quantum mechanics, Werner Heisenberg, once remarked, "We have to remember that what we observe is not nature herself, but nature exposed to our method of questioning." It can be nearly impossible to measure any human quality without altering the participant, and testing effects refer to any potential confound that occurs merely because the manner in which the participant was assessed. This is particularly a problem for studies involving repeated measurement, which is why testing effects have been variously referred to as progressive errors or carryover effects. When outcome measures are based on judgment raters or self-report measures, there is a heightened potential for testing effects. For example, at pretreatment a patient may provide a self-report assessment that refers to a high degree of likelihood of depression. The act of merely completing the assessment may provide some degree of abreaction that alleviates depression, and at post-treatment the patient may report decreased depression, even if the cause of the decrease was the testing device and not the treatment itself. Thus, self-report ratings may be biased

due to introspection. Additionally, various performance-based tests, whether a cardiac stress test or an intelligence test, are prone to a special type of testing threat, namely practice effects; that is, improvements over time may be due to increased familiarity or growth resulting from prior testing. In contrast, when prior testing depletes or diminishes physical or mental resources, declined performance may be the result of fatigue effects. Physically invasive procedures may also cause testing effects, for example, by alleviating pain or causing physical deterioration; thus testing effects can be either positive or negative. To combat this threat, control groups are generally used in between-group designs and multiple control phases in single subject designs, allowing the researcher to see testing threats in absence of the treatment.

Instrumentation

An instrumentation threat occurs when an observed effect might be due in part to inconsistencies in the testing device, raters, judges, or other instrumentation devices. This threat may occur when testing instruments are not standardized across groups or phases, such as non standardization of glucometers. Treatment effects could be exaggerated if the study draws upon inaccurate instruments for measurement of the outcome. To combat this threat, researchers should have quality-control standards in place, documenting the measurement equivalence of instruments across patients, groups of patients, or phases. Additionally, repeating phases in a single subject design can facilitate more confidence that the results are valid and do not contain measurement error.

Withdrawal Reactions

There are three central reasons why outcomes may worsen in response to the removal of a treatment [35]. First, original symptoms can reappear, often called relapse. Second, psychological factors or expectancy effects can cause the outcomes that are expected. Third, the withdrawal of some medications can cause rebound effects, aggravating symptoms beyond their original level, and although withdrawal reactions are frequently neglected, they can lead to an overestimation of treatment effectiveness. Many medications cause some degree of tolerance; that is, through feedback mechanisms the body regulates its own systems to compensate for actions caused by a medication. For example, in response to long-term use of synthetic steroids, the body compensates by producing fewer natural steroids, or engaging in other compensatory mechanisms. When a medication is then discontinued or substantially decreased, the body may have a diminished capacity for using its own natural resources, which can lead to symptom increases. Benzodiazepines, for example, are often used to treat symptoms of negative affectivity because they facilitate GABA transmission, producing a sedating effect. However, over time the body compensates for the medication by downregulating receptors for

GABA, minimizing the effects of the medication. Because the body compensates by dampening its own mechanisms for producing sedation, the abrupt withdrawal of a benzodiazepine would likely lead to a marked increase in anxiety, especially in comparison to the initial symptoms. Withdrawal reactions are common for various types of sedatives, stimulants, antidepressants, and antihypertensives [36]. Furthermore, there is considerable variability across individuals. Withdrawal reactions can pose problems for evaluating the internal validity of between-group and single subject designs. In between-group designs, often before beginning the study trial, patients go through a washout period in which all medications are withdrawn. Sometimes this washout phase is also used to measure initial symptoms; however, such an approach is problematic because symptoms during the washout phase would be exaggerated due to withdrawal reactions. If study outcomes are evaluated against baseline data collected during a washout phase, results will overestimate treatment effectiveness or efficacy. Within single subject designs, this problem is particularly important, especially if a medication is repeatedly compared to a placebo (e.g. A-B- A-B-A-B-A-B). If withdrawal reactions occur during the placebo phases, results would overestimate the benefits of the medication. Notably, withdrawal reactions dissipate overtime, so the solution to this problem is to ensure than non-treatment phases are lengthy enough to allow for symptoms to stabilize after withdrawal reactions dissipate. Unfortunately, physicians and researchers have failed to heed this threat, often using brief phases for studies involving stimulants [37, 38].

Social-Cognitive

Social-cognitive threats refer to the ways in which processing of social situations can potentially bias results. Examples include diffusion effects, compensatory rivalry, patient reactance, and self-fulfilling prophecies. Diffusion effects refer to any instance where components of an intervention inadvertently spread across groups or phases. In a between-group design, this could occur when patients in the control condition learn about a treatment option (e.g., exercise) and begin incorporating it into their own lives, with the consequences of reducing the differences between the conditions. For a single subject design, this may occur if a patient continues to self-administer a particular treatment during a non-treatment phase. To minimize diffusion threats, the researcher should emphasize to patients the importance of following protocols, provide incentives for following protocols, and use fidelity checks to monitor adherence to the protocol.

Compensatory rivalry occurs when patients increase motivation in a control condition to document their own personal strength or impress the researcher. This threat can occur in a between-group design when patients are aware they have been assigned to a control condition or in a single subject design during a baseline or non-preferred treatment phase. The researcher can deal with this threat by using the tactics for managing diffusion effects and also by encouraging patients to act as they typically do act, neither increasing nor decreasing their motivation.

In addition to improving their performance in a control condition, the results may also underestimate true effects if patients decrease their motivation in a treatment condition (i.e., patient reactance). Patient reactance can occur when participation is non-voluntary or when treatments are uncomfortable, time-consuming, or aversive. Although this limitation can occur in both between-group and single subject trials, the benefit of the single subject design is that a more individualized treatment plan can be implemented. Single subject studies have been shown to improve both treatment fidelity and outcome [21].

Further, self-fulfilling prophecies occur when patients' or researchers' expectations lead them to bring about the expected result. Often, self-fulfilling prophecies are discussed within the context of placebo or allegiance effects. Placebo effects occur when an intervention works solely or in part because patients expect it to work. Placebo effects have been most widely documented within the context of pharmaceutical research, but placebo effects can occur within the context of any type of intervention, from behavioral programs to cardiac surgery. It has been shown that placebo effects improved the outcomes in approximately 75% of biomedical studies [39, 40]. Similarly, allegiance effects occur when researchers' biases and expectations lead to more desirable results for a favored treatment. To guard against these threats, control conditions are often used. In single-blind (single-masked) procedures, the patient is unaware of the assignment, and in double-blind (double-masked) procedures, the patients and researchers administering the treatment are unaware of the assigned conditions. However, these methods of combatting expectancy effects have limitations. Even in double-blind (double-masked) randomized controlled trials, approximately 75% of patients and researchers are often accurate in guessing whether a placebo or actual treatment was being used [41]. Additionally, in a meta-analysis of antidepressants, McKay, Imel, and Wampold [42] found that allegiance effects actually account for more variance in outcomes than treatment. Further, merely using a "placebo" cannot control for all possible placebo effects. For example, many pharmaceutical studies use "inert" placebos, such as sugar pills or empty capsules, which have no major physiological effects and do not produce side effects. In contrast, "active" placebos can be chosen that produce mild physiological effects, such as increased autonomic arousal. Because active placebos are more difficult to distinguish from actual treatments, they produce placebo effects that are substantially larger [43]. To the extent that studies use weak placebo conditions, they will overestimate the efficacy of treatments, a disconcerting finding, given the high frequency of inert placebo use in randomized controlled trials.

Residual Confounding

To address threats to internal validity, researchers will often statistically or methodologically control for confounding variables. For example, in a randomized controlled trial, despite random assignment, the two groups of patients may differ

slightly in terms of initial symptoms, particularly if the sample size is small. Because this threatens internal validity, the researcher could statistically control for initial differences in symptoms. However, the ability to control for a confounding variable is only as strong as the researcher's ability to measure the variable. When a researcher fails to completely control for a third variable as a result of poor measurement, some portion of the confounding effect remains, known as residual confounding. Residual confounding has been frequently documented in epidemiological studies, where researchers face the difficulty of determining the relationship between two variables by partialling out the effects of various confounds. Attempts to statistically control for confounds are also common in between-group designs, specifically to control for baseline individual differences across groups. However, the threat to internal validity will remain if the confounding variables are poorly measured. Sometimes researchers will methodologically control for confounds; that is, rather than statistically controlling for differences in socioeconomic status and age, for example, exclusion criteria are used to ensure that patients are relatively homogenous. The extent to which patients are similar on confounding characteristics is the degree that those confounds will be controlled. Again, however, the ability to methodologically control for threats is only as strong as the quality of the measures used for excluding patients.

Threats to External Validity

Internal validity refers to the extent to which the researcher can infer causality between the independent and dependent variable. In contrast, external validity refers to the strength of results generalizing to other contexts. Most often, studies are conducted to produce generalizable knowledge; that is, whether the results of a study can be applied to similar cases and settings. Like internal validity, support for external validity is best viewed along a continuum. Typically, between-group studies are considered to have better external validity than single subject designs, but there are several techniques for countering this limitation [15, 23, 24, 31, 32]. The following sections describe how external validity differs across several contextual variables (Table 3.2).

Generalizability Across Subjects

An important consideration in evaluating the results of a study is whether the intervention will be similarly effective for different patient populations. This includes whether the results are similar across demographic groups based on age, sex, gender, race, ethnicity, socioeconomic status, among others. Also, researchers should consider whether results would be similar across individuals with different diagnostic characteristics, such as differences in onset, severity, allele frequency, disorder classification, or type of symptoms. Researchers may also be interested in whether results will generalize to patients with different, but related, diagnoses. Often, between-group studies are considered to have superior external validity across this

Table 3.2 External validity across Contextual Dimensions

Dimension	Description
Subjects	Results may differ across patients with different demographic characteristics, symptoms, or diagnoses.
Physicians/practitioners	Practitioner training, skill, experience, and fit may moderate results.
Settings	Results may be impacted by treatment handled in different locations or centers, along with implementation outside the research context.
Time	Results may vary depending on the time of day of the implementation, duration of the study, and historical context.
Outcomes	The results of a study depend on the manner in which outcomes defining success are quantified.
Treatment interactions	The effectiveness of a treatment may vary substantially, depending on potential interactions with concomitant interventions.

dimension because results are averaged across (i.e., summed over) a large number of patients [44]. However, as previously discussed, group means will not be predictive for all patients and demographic groups [15, 31]. When sample sizes are large enough for adequate power, a consideration of subgroup analyses is appropriate to examine whether the effectiveness of treatment is moderated by key demographic variables.

The ability to produce results that will generalize across patients is often considered a key limitation of single subject design studies. In only using one patient, it may be difficult to determine how the treatment would affect others. There are two methods for addressing this limitation: (1) the use of a prototypical patient or participant. This approach can be used to document that a treatment will work for a typical patient case; and (2) replication across a series of patients or participants. If a researcher can demonstrate that a treatment is similarly effective across a handful of diverse patients, practitioners can be more confident that the results will generalize to patients with other characteristics. Whereas the between-group design researcher merely attempts to average individual differences in treatment outcome, the single subject design researcher aims to exercise experimental control over treatment outcomes, modifying an intervention until the desired level of success is obtained. In this regard, single subject design studies may report on innovative techniques for obtaining desired outcomes for patients who might not respond to a generic intervention implemented in a between-group design.

Generalizability Across Physicians or Practitioners

The degree to which results vary across physicians or practitioners likely depends on the domain of research. For behavioral interventions, such as psychotherapy,

or performance-based interventions, such as surgery, the physician or practitioner plays a more important role than when treatment is self-administered by the patient, such as with medication. Of course, even with medication, the physician or practitioner can play an important role in moderating results [42]. In a single subject design, when it is important that results generalize across different physicians or practitioners, it may be useful to draw upon the multiple baseline design, extending the intervention to different physicians or practitioners one at a time.

Generalizability Across Settings

The setting in which an intervention is implemented can play an important role in the generalizability of results. Generalizability across settings is related to other contextual variables because different treatment centers have different patient populations and types of practitioners. Additionally, due to priming effects, the power of an intervention can also depend on contextual cues. Interestingly, when a medication is repeatedly taken within the same environment, the human body becomes primed to downregulate the response to the medication. In a novel environment, such cues are absent, so priming does not occur, and the medication may have a stronger impact, evidenced by the frequent overdose rates in individuals who abuse drugs when placed in novel environments [45]. Thus, researchers should keep in mind that interventions may have a more potent effect in novel environments.

Finally, it should also be considered whether similar results would be obtained in a non-research setting. A research setting is unique in that there is a greater presence of social-cognitive variables, such as diffusion effects, compensatory rivalry, patient reactance, and self-fulfilling prophecies, including placebo and allegiance effects. To the extent that these factors differ across settings or practitioners, the generalizability of results will be affected.

Generalizability Across Time

There are three ways in which results may vary due to temporal variables. At the simplest level, the researcher must consider whether the time of day will play a role in the results. This threat is particularly critical when medication or other interventions act only for a few hours, when outcomes may be affected by metabolic activity, or when the setting (e.g., home, school, or work) can affect outcomes. Although between-group designs may be relatively restricted in terms of design constraints, the single subject design affords important opportunities for handling this threat. Through the use of a multiple-baseline design, the researcher can examine whether the intervention varies in effectiveness throughout the day and potentially adjust the intervention accordingly. Additionally, it should be considered whether an effectiveness or efficacy of the intervention varies as a function of the duration of the study, and specifically when the final outcome measure is obtained. Whereas one treatment may outperform another in the short-term, it may prove inferior in the long-run. Finally, it should be noted that any study is conducted within a historical

context, and the intervention that is most effective today may not be in the future. The evolving nature of science assures that new and better treatments will continuously develop.

Generalizability Across Outcomes

Results may vary depending on the particular outcome measure that is used. This threat is important to consider because any particular intervention may have its own strengths and weaknesses. Convincing evidence for an intervention's external validity would come from evidence showing that the intervention is effective across multiple relevant outcomes. In this regard, single subject designs may have a slight advantage. Specifically, if an intervention only improves scores on one outcome measure, the intervention can be repeatedly altered until criterion levels are obtained on all relevant outcome measures.

Generalizability Across Treatment Interactions

Researchers need to consider how the results will vary when an intervention is implemented within the context of a treatment for other conditions. Many randomized controlled pharmaceutical trials examine treatments using only a single medication. However, in practice-based medicine, polypharmacology is common. Given the number of deaths and side-effects caused by drug-drug interactions, the generalizability of treatment outcomes in the context of other interventions can be difficult to predict [36]. Because single subject designs afford possibilities for monitoring patients more closely, they may prove useful in addressing this concern. Furthermore, single subject designs have been shown to be useful in reducing side effects and increasing treatment adherence [21].

Summary

This chapter highlighted the historical and contemporary foundations of research methodology as it applies to biomedicine and single subject research. Emphasis was placed on the strengths and weaknesses of single subject and between-subject designs. Although the single subject design affords a number of strengths, it has historically been overlooked in favor of between-group designs, in part due to statistical developments that catalyzed their use. Nonetheless, single subject designs can indeed play an important role in biomedical research and practice, particularly as it applies to internal validity. Despite the underutilization of the single subject design due to external validity concerns, more contemporary methodological approaches exist for overcoming these limitations, permitting the single subject design to play a more valuable role in biomedical research and practice.

References

1. Janosky JE, Leininger SL, Hoerger M. The use of single-subject methodology for research reported in biomedical journals. White Paper, Central Michigan University, 2009.
2. Forbes J. The British and Foreign Medical Review or Quarterly Journal of Practical Medicine and Surgery. 1840; 10(3).
3. Broca P. Remarks on the seat of the faculty of articulate language followed by an observation of aphemia. In G von Bonin (Ed.), *Some papers on the cerebral cortex* (pp. 49–72). Springfield, IL: C.C. Thomas, 1861/1960.
4. Fechner T. Elements of psychophysics. In H Langfeld (Ed.), *The classical psychologists* (pp. 562–572). Boston, MA: Houghton Mifflin, 1860/1912.
5. Ebbinghaus H. *Memory: A contribution to experimental psychology*, (HA Ruger, CE Bussenius, Trans.). New York: Columbia University Press, 1885/1913.
6. Pavlov IP. *Lectures on conditioned reflexes*. New York: International Publishers, 1928/1963.
7. Jay V. The legacy of Jean-Martin Charcot. *Archives of Pathology and Laboratory Medicine*. 2000; 124: 10–11.
8. Darwin C. *On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life*. London: John Murray, 1859.
9. Stilson DW. *Probability and statistics in psychological research and theory*. San Francisco, CA: Holden-Day, 1966.
10. Boring E. *A History of psychology* (2nd ed.). New York: Appleton-Century-Crofts, 1950.
11. Box J. Guinness, Gosset, Fisher, and small samples. *Statistical Science*. 1987; 2: 45–52.
12. Fisher RA. *Statistical methods for research workers*. Edinburgh: Oliver & Boyd, 1925.
13. Boring E. The nature and history of experimental control. *American Journal of Psychology*. 1954; 67: 573–589.
14. Dukes W. N = 1. *Psychological Bulletin*. 1965; 64: 74–79.
15. Barlow D, Hersen M. *Single case experimental designs: Strategies for studying behavior change* (2nd ed.). New York: Pergamon, 1984.
16. Healy D. *Let them eat Prozac*. New York: New York University Press, 2004.
17. Skinner BF. On the conditions for elicitation of certain eating reflexes. *Proceedings of the National Academy of Sciences*. 1930; 16: 433–438.
18. Skinner BF. *Science and human behavior*. New York: Macmillan, 1953.
19. Skinner BF. Behaviorism at fifty. *Science*. 1963; 140: 951–958.
20. Catania AC. *Learning: Interim* (4th ed.) New York: Sloan, 2007.
21. Avins AL, Bent S, Neuhaus JM. Use of an embedded N-of-1 trial to improve adherence and increase information from a clinical study. *Contemporary Clinical Trials*. 2005; 26: 397–401.
22. Institute of Medicine. Committee on strategies for small-number-participant clinical research trials, 2001.
23. Janosky JE. Use of the single subject design for practice based primary care research. *Postgraduate Medical Journal*. 2005; 81: 549–551.
24. Rapoff M, Stark L. Editorial: Journal of Pediatric Psychology statement of purpose: Section on single-subject studies. *Journal of Pediatric Psychology*. 2008; 33: 16–21.
25. Hume D. *An enquiry concerning human understanding*. New York: Bobbs-Merril Co.,1784/1984.
26. Janosky JE. Statistical testing alone and estimation plus testing: Reporting study outcomes in biomedical journals. *Statistics and Probability Letters*. 2008; 78: 2327–2331.
27. Hill AB. The environment and disease: Association or causation? *Proceedings of the Royal Society of Medicine*. 1965; 58: 295–300.
28. Susser M. What is a cause and how do we know one? A grammar for pragmatic epidemiology. *American Journal of Epidemiology*. 1991; 133: 635–648.
29. Weed D. Causal and preventive inference. In P Greenwald, B Kramer, D Weed (Eds.), *Cancer prevention and control* (pp. 285–302). New York: Marcel Dekker, 1995.

30. Weed D, Gorelic L. The practice of causal inference in cancer epidemiology. *Cancer Epidemiology, Biomarkers & Prevention*. 1996; 5: 303–311.
31. Kazdin A. *Single-case research designs: Methods for clinical and applied settings*. New York: Oxford University Press, 1982.
32. Richards S, Taylor R, Ramasamy R, Richards R. *Single subject research: Applications in educational and clinical settings*. Belmont, CA: Wadsworth, 1999.
33. Trochim W, Donnelly JP. *The research methods knowledge base* (3rd ed.). Mason, OH: Thomson Publishing, 2007.
34. Dimidjian S, Hollon S, Dobson K, et al. Randomized trial of behavioral activation, cognitive therapy, and antidepressant medication in the acute treatment of adults with major depression. *Journal of Consulting and Clinical Psychology*. 2006; 74: 658–670.
35. Breggin P, Cohen D. *Your drug may be your problem*. New York: Perseus Books, 1999.
36. Reiss S, Aman M. *Psychotropic medications & developmental disabilities: The international consensus handbook.* Columbus, OH: The Ohio State University, Nisonger Center Publisher, 1997.
37. Johnson C, Handen B, Lubetsky M, Sacco K. Efficacy of methylphenidate and behavioral intervention on classroom behavior in children with ADHD and mental retardation. *Behavior Modification*. 1994; 18: 470–487.
38. Nikles C, Mitchell G, Del Mar C, Clavarino A, McNairn N. An n-of-1 trial service in clinical practice: Testing the effectiveness of stimulants for attention-deficit/hyperactivity disorder. *Pediatrics*. 2006; 117: 2040–2046.
39. Benson H, Friedman R. Harnessing the power of the placebo effect and renaming it "remembered wellness". *Annual Review of Medicine*. 1996; 47: 193–199.
40. Guess H, Kleinman A, Kusek J, Engel L. S *cience of the placebo: Toward an interdisciplinary research agenda*. London: BMJ Books, 2002.
41. Vitiello B, Davis M, Greenhill L, Pine D. Blindness of clinical evaluators, parents, and children in a placebo-controlled trial of fluvoxamine. *Journal of Child and Adolescent Psychopharmacology*. 2006; 16: 219–225.
42. McKay K, Imel Z, Wampold B. Psychiatrist effects in the psychopharmacological treatment of depression. *Journal of Affective Disorders*. 2006; 92: 287–290.
43. Kirsch I. Are drug and placebo effects in depression additive? *Biological Psychiatry*. 2000; 47: 733–735.
44. Newcombe R. Should the single subject design be regarded as a valid alternative to the randomised controlled trial? *Postgraduate Medical Journal*. 2005; 81: 546–547.
45. Madden GJ. A behavioral-economics primer. In. WK Bickel, R Vuchinich (Eds.), *Reframing health behavior change with behavioral economics* (pp. 3–26). Mahwah, NJ: Lawrence Erlbaum & Associates, 2000.