2016

Potential Problems in Objectivist Research

Arthur P. Sullivan
Touro College, asullivan75004@gmail.com

John A. Sullivan

Chapter 17

POTENTIAL PROBLEMS IN OBJECTIVIST RESEARCH

Arthur P. Sullivan • John A. Sullivan

It can be disheartening for a music therapist to experience the effects of therapy in the clinic, yet fail to find significant effects when a research trial is performed. Many real effects remain unsubstantiated, even in apparently well-designed studies. This can divert music therapy in unproductive directions. Further, when research is in preliminary stages, any finding that indicates the investigation is on the right track is valuable (e.g., Nayak, Wheeler, Shiflett, & Agostinelli, 2000). But as findings accumulate and are surveyed and reviewed, as they have been in music therapy research (e.g., a number of Cochrane reviews that have been done on music and music therapy, www.cochranelibrary.com), more is at stake for the researchers and ultimately for the patients and clients who will be affected by what is learned or could have been learned. For these reasons, music therapy researchers are encouraged to seek more than just a match among the parts of the research project. Seeking an excellent or at least a best-possible match is proposed.

Pursuing the anatomy of a best possible match starts by identifying the problems that remain after a reasonable match has been made. The breadth of coverage of objectivist research in music therapy presented in this volume assures the careful reader the procedural knowledge required to create a good research design. Since music therapy research has reached a maturity where vulnerability to pitfalls has become important, a competent overview must attempt to blend the several parts of the research process while focusing on any potential pitfalls and how they can be remedied or their damage attenuated.

Potential Problems

Some of the problems that can occur will be raised here. Possible solutions will be presented in the next section.

Failing to detect a real effect can occur when the researcher has not paid attention to sometimes enormous ***Type II error*** probabilities. Validating the success of a procedure in the clinic can keep researchers focused on ***Type I error*** alone. Each researcher wishes to deliver therapeutic effectiveness which cannot reasonably be attributed to chance, so the goal is an error probability of less than .05; better still less than .01 or .001. These low probabilities provide increasing likelihood that findings are not measurement artifacts and that the client improvements observed are actually present.

Yet the related question is typically not addressed and often not even considered. What is the

likelihood that there was real client improvement which the research failed to detect? This is the Type II error question. In many research projects, this probability is between 40% and 60%. Because this probability was not calculated, the researcher proceeded with an experiment wherein the effect was quite unlikely to be found and then discarded the procedure being tested as useless because the effect was not found, or was *not significant*.

Correcting this problem is not so easy as adjusting a Type I error probability, but Type II error probability can be reduced. Indeed, it often must be reduced if there is to be any hope of finding the effect.

Type II error neglect is not the only potential pitfall. It can also be frustrating for a music therapist who has done some preliminary research and found client improvement to then be *unable to demonstrate a causal link* between client improvement and music therapy. This often happens in research projects because of the small number of participants available, but more typically because of the therapist's being unable or unwilling—for compassionate, practical, or ethical reasons—to construct a true, randomly assigned control group. There are fixes: Causality can be demonstrated without using a randomized control group, for example, with a regression–discontinuity design as discussed below.

Regression to the mean, either unnoticed or inadequately addressed, is another potential problem. The most obvious scenario plays out frequently because treatments are created for extreme groups, specifically for clients with problems. Posttesting is doomed to show improvement because of regression to the mean, and true effects can be lost among the artifactual ones.

The situation is greatly exacerbated when clients are selected for the study by administering a selection test at random. The probability that randomly administered tests correctly find the clients in need of therapy is quite surprisingly low, even when the percentage of suitable clients in the sampled population is relatively high and the test has reasonably good psychometric properties. For example, in a patient population of 100 where 10 are in reality suited for music therapy, a screening test with .90 sensitivity and .90 selectivity can be expected to identify 9 of those 10 patients. But it will also identify 8 patients incorrectly, giving the researcher a treatment group of 17 in which almost half are unsuitable. Getting a more accurate screening test barely helps: Fewer patients will be identified, but still only around 50% of them will actually be suitable. These inappropriately selected clients can sharply increase the posttest regression effect, particularly in otherwise suitable, even posttest-only designs.

Researchers who are not also statisticians can underestimate the damage done by *violating statistical assumptions*. The statistical assumptions are made to allow solving of the mathematical equations that underlie the statistical procedure. Even small violations can do unexpectedly large damage and cause misleading results, for example, finding random fluctuations to be significant or failing to find an effect that is actually present. These problems are often correctible by minor changes in the statistical analysis or by changes in the research design.

Finding a statistically significant result is invariably encouraging. But pursuing it without estimating the ***effect size*** or confusing the level of significance with the size of the effect may leave the therapist expending far too much effort in therapy for very little gain for the client. A businessperson would take a cost–benefit approach to the situation to determine whether the effect found was worth pursuing. While no researcher wants to abandon a significant finding, perhaps a hint of the cost–benefit approach is useful.

Cautions are not all provided by statisticians. Some decision theorists (Kahneman, 2011) report a tendency for people in general to *answer easier questions than those that were asked.* A school principal trying to inhibit behavior asks himself the question, "What punishment does this student deserve for what he did?" That is not precisely the right question, but may be sufficiently close. However, the principal might inadvertently answer that question, "How much does what the student did make my job more difficult?" Researchers, ever oriented to the actual questions they wish to ask and answer, can still pursue easier questions and inadvertently predicate the answers they find for the easier questions as answers to the real research questions. This can be the unnoticed by-product of the piecemeal topography of the research process, or it can occur unobtrusively when operational definitions are being decided.

These are considerations that rise to the level of concern in music therapy research and, in fact, in all research. This overview undertakes to set them in useful context and point to possible solutions.

Recognizing and Ameliorating Problems

Although insight about any part of the research project can occur to the researcher at any time, the process has a structured flow. Specifying the flow is not intended to impede the researcher's thought process, but to organize the results of it. A new insight typically causes an important change to one part of the project, but also causes secondary changes in other parts. Keeping the overall picture in focus helps to adjust all parts of the process for an improvement made in any one part.

In general, the sequence of the superordinate elements that form the spine of the research project is: (a) research questions, (b) creation of research design, (c) adoption of measurement instruments, (d) statistical analysis, and (e) translation to clinical practice. The problems that plague the researcher, often enough without being noticed, can appear differently in different parts of the research process. Some problems occur exclusively at certain points in a research project. The following sections organize the problems within the part of the research process where they are likely to occur, and managing them is addressed in that context.

In the Research Questions

The research question begins with the researcher's curiosity and contains comparisons. When the comparison is not explicit, it is there anyway, implicitly, even in pheno meno logical research. Although there might seem to be no comparison involved when determining if, when, to what degree, and how often an event occurs, or whether an event is predictable, comparison exists, to a standard or at the very least to a background against which the observation is made. Total lack of contrast prevents a figure from being discernible from the ground. The first formalization of the researcher's curiosity is to make the comparison clear.

Whether the comparison is to a standard norm or metric, between groups of persons, or between persons and themselves at a different time will trigger a cascade of decisions. The nature of the comparison narrows the possible choices of research design, which then affects the choices in every

part of the research process.

This book includes a taxonomy of research questions (see Figures 1 and 2 in Chapter 19, Introduction to SPSS), and it is very probable that the reader will find the form of the desired research question on this list. After selecting the most similar question, the reader can then use the decision tree there to see which statistical methods and associated research designs are useful for that type of comparison.

Problem: Answering an Easier Question. It is at this point, while specifying the research questions, that the researcher is encouraged to become watchful about inadvertently substituting an easier question. It is easier to determine whether patients enjoyed the therapy than it is to determine whether it attenuated their dysphoria. Measures of anxiety and depression, for example, can devolve into measurements of enjoyment of the sessions or appreciation of the therapist if the research elements are not selected with a view to preventing such substitutions.

In the Creation of the Research Design

Choice of research design is often merged with considerations of the statistical analyses to be performed. This is a useful approach and is reflected in the structure of the decision tree from Chapter 19 in this book, referred to above, which omits separate mention of the research design because the statistics are often closely wedded to the specific designs. In overview, however, the quality of the match between statistics and research question can be inspected for possible improvements that can assist in avoiding research problems.

Useful research designs include those to which chapters of this book are devoted: case study; AB, ABA, ABAB, and other withdrawal designs; multiple baseline, changing criterion, and multiple treatment designs; survey research; longitudinal designs; one-sample designs; static group comparisons; parallel group designs; crossover designs; and factorial designs. Other designs that may be useful are time series/repeated measures, counterbalanced, Solomon four group, and Latin square; information about these designs is also found in this book.

The connection between design and statistical analysis is evident. Some of the above designs (e.g., case study, withdrawal, multiple baseline, changing criterion, and multiple treatment) do not require statistics, although they may be used in some cases. Estimation of population parameters requires descriptive statistics and graphics only. Parallel group, crossover, and factorial designs, among others, call for *t* tests and various ANOVAs, and ANCOVAs when covariates are present. Several designs, such as repeated measures and Solomon four group, are often overlooked even when the researcher has designed a near equivalent (Creswell, 2012). These designs are useful alternatives that change procedure only but do not disrupt the connection between research question and statistical analysis. Using these alternative designs can facilitate the research effort and avoid some of the problems under consideration.

Problem: Attributing Causality Without a Control Group. The ***regression-discontinuity***

design is used when random assignment to a control group is not practical, is not desirable, or is unethical. A review of the capabilities and limitations of the regression-discontinuity design is presented by Imbens and Lemieux (2007). This design can be used when the participants are screened for severity of symptoms so that those with the most severe symptoms, for example, in the worst pain, will receive the treatment first, while those with less severe symptoms will be placed on a wait list and receive treatment later. The screening has a cutoff point for severity. Those very near the cutoff point on either side of it have very similar symptom severity, and the probability is high that they fell on one or other side of the cutoff by chance arising from ***measurement error.*** The essence of this design is in comparing the near-the-cutoff participants who were barely included for treatment (those just above the cutoff) to those who were barely excluded from treatment (just below the cutoff point).

Statistically, this subgroup of patients near the cutoff can be treated as if they were randomly assigned to treatment or control groups, and the ordinary least-squares procedures such as ANOVA, ANCOVA, and the whole range of linear and multiple regression methods may be used. The conclusions are causal provided that the researcher did not disturb the cutoff process, for example, by compassionate inclusions of patients in the treatment group who fell below the cutoff but whom the researcher felt should get treatment without delay. The group assignment to experimental and control was not perfectly random but rather *as good as random* and further disturbance of the process may invalidate it.

The drawback to this design is that only a fraction of the patients treated can be viewed as near the cutoff point and included in the study, so a number of cycles may have to be run before a sufficient sample size is attained.

Problem: Failing to Detect a Real Effect—Too Few Participants. Not finding an effect that is actually present is discussed under the headings of *Type II error* or *low power* when it is discussed or noticed at all. Researchers can think that if they have a large enough sample size, this is not a concern, and, to a point, they are correct. But considering the time and effort spent in providing music therapy treatment, the number of participants a researcher can include in a reasonable period of time is often well below what the researcher would desire. In this event, specialized designs might be worth considering.

One method for increasing the power of research with few participants is through the use of a repeated measures design (Ellis, 1999). The researcher using this design measures client change repeatedly, often beginning with a pretest. In a study where the music therapy procedures are intended to reduce client distress, the pretest measurement establishes the baseline level of distress for all participants. The researcher then repeats the measurement, perhaps five times over 15 therapy sessions. This has the effect of increasing the power of the experiment to the level of one with almost five times as many participants, thus greatly reducing the likelihood of missing a real improvement in the clients if an improvement actually happens.

Where very few clients are available, a ***Latin square design*** can be adopted under certain conditions. Latin squares have been popularized among the general public in the Sudoku puzzles. Like these puzzles in which a number must appear only once in each row or column, each research subject must appear once and only once in each condition and level. Using this design, a music therapy researcher wishing to study, for example, the effects of music therapy and anxiety medication on patient distress could use a Latin square design requiring only three clients yet have the statistical

power of a three-way ANOVA design requiring at least 15 times as many client participants.

To do this, the researcher would define three levels of music therapy (perhaps number of sessions or variations in the therapy procedure) and three levels of prescription anxiolytics (perhaps zero plus two other dosage levels). Each client's distress, the dependent measure, is assessed in each of the three conditions he or she experienced.

There are restrictions. Factor levels must be randomly assignable. Thus levels of patient pain, illness, or gender cannot be used, but levels of exposure to treatment, therapist, or procedures can be used. Like the rows, columns, and numbers in Sudoku puzzles, the number of clients, levels of treatment, and levels of medication must all be equal. Also like Sudoku, each client will appear once only in each row (level of therapy) and column (medication dose), and these must be randomly assigned: That is, for the research, each of the three participants is assigned to a level of therapy and medication three times, at random. The excellent leverage in this procedure can be attractive to patients who desire to make a personal contribution to scientific knowledge and prefer participating in this type of study rather than in a conventionally designed, large-scale study where they are only one of very many participants.

Note that the design has additional advantages beyond requiring so few participants. The conclusions are causal, with effective control of extraneous error sources. Conclusions about treatment versus no treatment can be made, as in the example here, by including a zero level in the treatment factor. The statistical analysis is ANOVA and can be done with SPSS, though the guidance of a statistician is suggested.

Problem: Failing to Detect a Real Effect—Treatment Is Lengthy or Patients Are Treated Sequentially. Obtaining a treatment effect in music therapy may take many sessions that extend over months. Additionally, the music therapist conducting the research can provide therapy to only a few clients at a time. Aggregating enough clients this way could take years, even with cooperating therapists including their patients in the study by using the experimental procedures.

In an instance like this where both time and number of clients are issues, *acceptance sampling* may be considered. Typically seen only in business and industry, the expense of providing treatment to each music therapy client amply motivates the straightforward adaptation of the procedure to music therapy research. The statistical analyses can be performed after every few clients have completed the treatment (or placebo treatment if they were assigned to the control group) as if the full planned number of participants had been reached. The experiment is halted when a significant result is obtained, or when the planned number of clients has completed it, whichever comes first. A significant result that will have a large effect size can emerge early, after relatively few clients have been treated, conserving resources. When no significant result occurs, it may become obvious that this will happen before all the planned clients have been treated. This is a disappointing outcome, but again, it will conserve resources for other research leads. Note that the need for random inclusion of additional clients for successive analyses is critical.

Problem: Failing to Detect a Real Effect—Variables Are Related Nonlinearly. The relationship between anxiety and performance is a well-studied example of this. As anxiety increases, the client's performance tends to improve, to a point. After that point, further increases in anxiety tend

to deteriorate performance. Anxiety is thus both positively and negatively related to performance, so an experimental procedure effectively reducing a client's anxiety will improve the performance of some of the clients and worsen the performance of others, yielding a net finding that reducing anxiety has not occurred or has little or no desirable effect.

This problem is usually addressed statistically. The data analysis should routinely include scatterplots of each ***independent variable*** with each dependent variable. If the shape of the scatter departs markedly from a roughly oval-shaped grouping, research results from statistical methods that assume linearity are being reduced by nonlinearity of relationship between the variables. This can be avoided by *curve fitting* (finding mathematical lines, curves, and surfaces that fit the data well and both elucidate and display its meaning) or corrected by *transforming* the nonlinear variables to near-linear to facilitate both analysis and interpretation or by using *separate statistical analyses* for different segments of the data range. In this example relating anxiety to performance, *separate statistical analysis* would involve merely performing separate analyses for the parts of the range in which the variables are positively related and the parts where they are negatively related. In the case of anxiety, really all three of these methods are viable candidates for fixing the problem.

One further note on scatterplots: If the scatter is roughly oval-shaped, but the oval is angled markedly different from 45 degrees, research results will be attenuated because of *restriction of range*. This is typically corrected by changing the metric of the variables or by increasing the sample size in the hope of increasing the range of values for the variable with the restricted range.

Problem: Regression to the Mean—Extreme Groups Are Being Studied. The very-worst-case scenario is when a researcher uses a measure to find the clients who need the treatment, provides the treatment to these clients, and then uses the same measure to detect treatment success. Measurement error, which is larger in extreme scores, will virtually guarantee that the researcher will find a significant client improvement, and possibly a quite large one. This research was doomed to seemingly succeed; how much of the improvement is regression to the mean artifact is unknown. This seriously compromised design is sometimes used when a researcher is working on a new method or an improvement on an existing method of therapy. The researcher tries it out on a small group of clients with difficult or treatment-resistant problems, and it seems to work because of regression artifact. The researcher may then spend scarce resources pursuing the development of a method that really does not work. When studying extreme groups, using a randomly assigned control group is essential. Analyzing the outcome data with ANCOVA, using the selection test as the covariate, further reduces error fluctuations and improves the likelihood of research findings that prove to be veridical.

Problem: Failing to Detect a Real Effect—Clients Are Found by Screening Tests. When researchers find study participants by random screening or by screening the entire population of a facility, 40%–60% of the participants found to be eligible are not actually eligible. Even if the screening test has excellent psychometric properties, this happens. The cause is in the baseline percentage of eligible clients in the group screened. The lower the percentage, the greater the number of ineligible clients who will be incorrectly identified. And the ineligible participants can prevent any true treatment effect from being detected by dilution.

One approach to avoiding this problem is to restrict screening to potential clients who have

already been identified in some other way as likely to be eligible. For example, instead of screening an entire facility to find eligible candidates, only those who have been found appropriate by clinical examination would be screened. This approach reduces inappropriate participants to statistically manageable levels.

Problem: Attributing Causality—Concerns About the Effect of the Pretest and Other Persistent Concerns About Validity Threats. Concerns about ***pretest sensitization*** affecting posttest outcomes independently of treatment is the most common of this class of problems, but other concerns about pretest effects or alternative causes for posttest outcomes can concern the researcher. The ***Solomon four group design*** allows the researcher to test for the presence of these flaws in the research outcome.

This design is actually a fully randomized pretest/posttest control group design together with a fully randomized posttest-only control group design combined. The multiple comparisons possible among the two pretests and four posttests with this design make the concerns that might otherwise have been study flaws directly testable and possibly correctable if present. This design has excellent statistical power and is very robust in the face of challenges to internal and external validity. It can, however, be expensive and difficult to administer because of the strict randomization and the number of groups. Consequently, weaker designs are typically selected.

In the Adoption of Measurement Instruments

Research questions typically permit a number of possible dependent variables, and the researcher chooses among them to find an optimal fit: one that most closely expresses the central intent of the research question, while remaining practical. For example, a researcher testing the clinical effectiveness of music therapy in reducing client distress might select depression, anxiety, or any other dysphoric state as the dependent variable, and in doing so, define *distress* for the study. The precise operational definition of the selected variable will be the method by which it is measured. In the instance of anxiety, the method could be the client's score on an existing research measurement, such as an existing test for severity of anxiety, or on a commercially available test of anxiety. It could also be predetermined observations and measurements made by the researchers, as in the example above on priming research. Or it could be a test authored by the researchers built to their own exact needs.

An example of the value of a good operational definition of the dependent variable concerns the priming effect (Bargh, Chen, & Burrows, 1996), in which students primed with stereotypes of older adults in an apparently unrelated word task walked significantly more slowly down the corridor to the next task than did controls. The contents of pre- and postinterviews and careful experimental design protected the conclusion from alternative explanations, and the study is a compelling demonstration of the aspect of priming known as the *Florida Effect*. Operationally defining *aging* as *slower walking speed* enabled the study.

Problems appear in the selection of variables and creating their operational definitions, as in other parts of the research process which merit researcher attention and adjusting. Operational

definitions work best if they are intuitive, compelling, and minimally technical, as in this example.

Problem: Answering an Easier Question—Selecting an Existing Test. Many tests exist in the research literature and are available through the author or commercial channels, and it is likely that a good enough test already exists for most clinical studies. Any selection will have advantages and disadvantages. The disadvantages that make the test a measure of something even slightly different from what the research question asks must be weighed carefully. If the match is not precise, the researcher might wish to consider a different mode of measuring the variable of interest or creating a test exactly suited to the research purpose.

Problem: Answering an Easier Question—Authoring or Customizing a Test for the Research Project. In creating a test instrument or observation procedure to operationally measure the variable of interest, the researcher might follow the following steps to validate a test for use in research: (a) specify the construct; (b) review the related literature; (c) specify the need for the new measure and how it differs from what is available; (d) specify the domain; (e) specify the structure (unidimensional scale, subscales, etc.); (f) create initial items; (g) conduct an initial pilot test on a convenient sample; (h) conduct a reliability analysis and repair problems; (i) perform an initial validity procedure using structural exploration with factor analysis; (j) perform a second validity procedure (the gold standard is the Multitrait-Multimethod Matrix [MTMMM; Campbell & Fiske, 1959; Cronbach & Meehl, 1955], but consider other appropriate demonstrations of validity suited to your measurement); (k) perform a second pilot test on a representative but small sample; and (l) when still working with small pilot samples and having improved/added/deleted items to improve psychometric properties (benchmark), hold on to the best reliability and validity estimates for your plans to validate the test on an appropriately sized validation and norming sample. Watch for deviations from these benchmark values by ongoing recalculations as you collect data.

These steps include iterations. If, for example, the statistical properties of the initial test items are not adequate, some rewriting is needed, followed by a trial of the new item set. The factor analysis should reveal a structure anticipated by the test author or one congruent with the researcher's purpose. It is important to select *display* of the correlation matrix determinant. Some programs, including some versions of SPSS, do not automatically display this quantity, even when it is exactly equal to zero, which happens frequently enough. A true solution with a zero determinant cannot exist because it includes division by zero, and the program may create potentially irrelevant output without warning.

Once discovered, the zero-determinant problem can usually be remedied, depending on its cause. One cause of a zero determinant is one item being a linear combination of other items. Two perfectly correlated items is an obvious case of this. Drop one of the items, since it adds no information. Less obvious is when two or more items exactly predict another item. This happens when scale scores or the total score of the test have inadvertently been included in the analysis. These scores are summations of items, which predict the total and scale scores perfectly. Eliminate these and rerun the analysis. Even less obvious is when some items predict other items perfectly in a given sample because of the nature of the content. Watch for lack of error variance or negative error variance and repair or remove the problematic items. If the problematic items are difficult to find, stepwise linear regression can be useful. These hints are not exhaustive, and researchers validating

their own tests will soon learn to check for many other causes when the determinant is exactly zero.

Problem: Failing to Detect a Real Effect—Scaling Level of Measure Too Low. While almost any type of measurement will lend itself to statistical analysis, higher levels of scaling will provide incrementally more statistical power. Select instruments at the highest scaling level the content permits, typically interval or ratio levels. When the researcher decides to author an instrument to meet the research needs more precisely, it is worthwhile to design it to provide measurements at the highest levels of scaling possible. Many variables in psychology and sociology are not demonstrably measured with interval-level scales, and treating them as interval scales when they are in reality ordinal-level introduces additional error. If a fully interval-level measure cannot be fitted to a research variable, a strong alternative that may require some help from a statistician is a Thurstone scaling. Ordinal processes in general can be measured at interval level or better by techniques like Thurstone scaling, particularly Case V scaling, which depends on the law of comparative judgment (Thurstone, 1927), an intrinsically ordinal process. Even lower-level scaling can be analyzed by powerful interval-level statistics: Nominal data can be converted to binary variables, which can then be used for some least squares statistics, such as regression, by employing a system such as dummy coding. Consulting with a statistician or someone with experience in psychometrics would also be advisable for help in dealing with this problem.

Problem: Failing to Detect a Real Effect—Inadequacy of the Dependent Variable. Some of the applications of music therapy can be viewed as influencing a developmental trajectory, rather than producing a fixed effect. Children who are developing normally but achieve developmental milestones more slowly than desired are exposed to interventions to accelerate the developmental process, thereby reducing distress and the sequelae of being routinely behind their peers. Complete development likely would have occurred anyway, and it is only the speed that was affected.

A typical outcome variable in the social sciences can be insensitive to this type of change. The application of music therapy may have measurable effects over time whereby the trajectory (i.e., the dependent variable) is significantly altered despite the final outcome remaining unchanged. Music therapy employed to bring relief to clients in the final stages of degenerative processes which will result in death would seem to be better studied by defining variables designed to detect this change in the trajectory of the degenerative process. In addition to the specific sensitivity of variables like this to the effect of music therapy as applied to certain clients, this method brings the added advantage of requiring very few research participants.

To apply this remedy, the researcher would define the dependent variable as a change in *slopes* of a line's tangent to the developmental curve at the testing point times. For example, despite medical interventions, dementia has an insidious degenerative course ultimately resulting in death. Interventions used to slow the rate of degeneration must look at variables that alter the course of the disease, but not necessarily the endpoint. The application of music therapy in dementia patients could slow the rate of disease progression without influencing survival time, a change reflected in change or rate or slope.

Alternatively, if the researcher hypothesizes that survival time is influenced, *Kaplan–Meier estimation* may be useful. These procedures are currently more widely used in the biological and

other-than-social sciences but may find useful application in music therapy. For researchers unfamiliar with recent developments in *latent curve analysis* and related procedures, consultation with a statistician is suggested.

In the Statistical Analysis

The use of particular statistical procedures to investigate research questions within various research designs allows considerably more creative leeway than usually may be supposed. The most conservative and widely accepted coupling of statistical procedures to designs and research questions has been well expounded in other chapters (e.g., Chapters 18 and 19). When a complex or creative statistical approach may be needed, a consultation with a statistician is suggested.

Table 1. Primary Statistical Tools

	Factor analysis
r, τ, ρ, Ï†	Correlations
f	Discriminant function
f	ANOVA, rANOVA, MANOVA, ANCOVA
z	*z*-test, percentiles
t	*t* test
Å·, β	Multiple regression, polynomial regression
μ, σ	Descriptives & estimation
z, t	Confidence intervals
Z	Linear programming
p	Simple probability
f	Repeated measures
χ2	Chi square
U	Mann–Whitney
R2 , D, η2, ω2	Size of effect, % variability accounted for
K-S	Kolmogorov–Smirnov goodness of fit
W	Wilcoxon signed-rank test
	Runs test for randomness

The researcher's basic statistical toolkit should include at least the tools listed in Table 1 (see

Chapter 18 for additional information). Each has underlying assumptions, depending upon its application, and violation of these assumptions can greatly increase the probability of an erroneous outcome. Before employing a statistical procedure, it is useful to check what ***assumptions*** must be made. Many procedures are *robust* with respect to assumption violations. Robust, however, is not a synonym for impervious. The greater the violation of assumptions, the greater the likelihood that the statistic will not perform properly.

Problem: Assumption Violations—Unequal Variances. Many statistics assume that the variances within different groups are equal, and each statistic has a margin of tolerance for violation of this assumption. ANOVA is one of the most tolerant. Rather than ignoring unequal variances and relying on robustness or using a *rule of thumb,* a simple statistical test such as Levene's or Bartlett's will determine if the violation is excessive. If the test is failed, Welch's ANOVA is an option. Alternatively, transforming data to *z* values will correct for the unequal variances, also termed *heteroscedascity,* and any non-normality at the same time. Before using a correction, be sure to plot the data and determine whether the cause could be incorrectly coded data, or, more importantly, if the existence of outliers reveals important information about the research question, procedures, or sample.

Problem: Effect Size. A significant result is not necessarily an important or useful one. Depending on the selection of alpha, the sample size, and the power of procedures, a certain number of significant results will occur by chance alone. These results are quite unlikely to replicate. Legitimately significant results, however, may still not be important. A significant change in therapeutic procedure yielding only a 1% improvement in outcome will have its main value in improving understanding of the procedure or pointing the researcher in a useful direction. The change in and of itself will hardly be worth the time and effort in clinical application. Resources are usually limited or scarce in research, so they need to be allocated to the most important gains or the most promising leads. To manage these concerns, a statistic that estimates the ***effect size*** is calculated whenever a significant result has been found, to give the researchers a grasp on the magnitude of the effect that has been found.

The size of the improvement is a useful indicator of importance. The size can be viewed as a mean change (How much improvement, on average, was there in clients' outcomes?) or a percent of variability (Of all the variability in the clients' outcomes, what percent is attributable to this change in therapy procedure?). It is simple to calculate, readily available in SPSS, and provides an essential context for evaluating any significant outcomes of research.

Problem: Failing to Detect a Real Effect—Low Statistical Power. The power of a statistical procedure is the probability that it will find a treatment effect, if one exists. The power depends primarily on sample size, alpha, and the size of the anticipated effect. The population mean and variance need to be known, but they can be estimated from a sample. The sample size and alpha are chosen by the researcher. The size of the anticipated effect can be derived theoretically (this therapy should alter client response to this many items and therefore lower the dysphoria score 10 points or more) or estimated from the sample in the pilot study. Thus all quantities are known, or estimable.

From these, the power can be calculated with a *t* test.

The power should be calculated before the research is begun, as soon as the quantities above can be estimated. There are two reasons for this. First, if the power was 42%, for example, it is doubtful that any researcher who knew this would undertake to do the research. There would be less than a 50/50 chance of detecting whether the therapy worked. Changes would have to be made if this research were to be done at all, but calculating the power in advance is the only way to know this. Software that assists in estimating power (e.g., PASS, G*POWER) could also be useful.

Second, the calculation can be used a different way. By the researcher's deciding that the minimum acceptable power is, for example, 90% (in effect deciding that there must be a 90% chance of finding the effect if it exists), the researcher can then calculate how large a sample size would be required to achieve this power. Knowing the resources of time, money, etc., required per client, the cost can then be estimated and a determination made whether the study is worth the cost.

Problem: Failing to Detect a Real Effect—Insensitive Dependent Measure. A dependent variable defining a fixed effect may be insensitive to a treatment that modifies a developmental process, as described above. Defining variables that quantify rate of change rather than magnitude of change either continuously or from one measurement point to another may find usefulness in music therapy research. The statistical power of repeated measures is attractive for music therapy research for reasons discussed above. Adding the statistical power of use of slope as a dependent variable might enable a more sophisticated examination of the effects of music therapy and a finer attuning of its use to client needs.

In Translation to Clinical Practice

Once the research has been carried out and the results are known, along with their effect sizes, the findings are reported in both statistical terms and in the parlance of scientist/researchers. The significant and important findings reside thenceforth in a filed report or a scientific journal article. In both places, it will be read but probably not acted on.

It is well established in research literature and academic discussion (Kahneman, 2011) that statistical arguments, however compelling, convince people cognitively but do not routinely influence their decisions. It is conjectured that readers acknowledge scientific findings while exempting themselves and those in contact with them from their applicability. However, it is also known to those studying availability heuristics that a single anecdotal tale, especially one with emotional connection, will influence decision-making and behavior, even if the anecdote is about a very rare instance.

Therefore, the communication of the research results, if they are to be incorporated into clinical practice by nonprofessional caregivers, includes more than reporting. It includes a separate task of communicating them to potential users in a manner that will affect their decision-making, that is, cause them to believe the research results so that they will be able to accept the findings and act on them. Ironies abound here. Research suggests that the most sophisticated findings might most effectively be translated to practice when framed as a case study. But case study is a research design thought by many to be a less sophisticated type of research than what may have led to the findings, an amusing

irony in reporting the results of research.

Conclusion

This chapter has outlined potential problems that may occur in the design of an objectivist research study and means of ameliorating them. Problems addressed in the chapter are divided into those that occur in the research questions, in the creation of the research design, in the adoption of measurement instruments, in the statistical analysis, and in translation to clinical practice.

References

Bargh, J. A., Chen, M., & Burrows, L. (1996). Automaticity of social behavior: Direct effects of trait construct and stereotype priming on action. *Journal of Personality and Social Psychology*, 71, 230–244. doi:10.1037//00223514.71.2.230

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait–multimethod matrix. *Psychological Bulletin, 56*, 81–105. doi:10.1037/h0046016

Creswell, J. (2012). *Educational research: Planning, conducting, and evaluating quantitative and qualitative research* (4th ed.). Boston, MA: Pearson.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*, 281–302. doi:10.1037/h0040957

Ellis, M. V. (1999). Repeated measures designs. *The Counseling Psychologist*, *27,* 552–578. doi:10.1177/0011000099274004

Imbens, G., & Lemieux, T. (2007). *Regression-discontinuity designs: A guide to practice*. Retrieved from http://www.nber.org/papers/w13039. Cambridge, MA: National Bureau of Economic Research. doi:10.3386/w13039

Kahneman, D. (2011). *Thinking, fast and slow.* New York, NY: Farrar, Straus, and Giroux. doi:10.1086/674372

Nayak, S., Wheeler, B., Shiflett, S., & Agostinelli, S. (2000). Effect of music therapy on mood and social interaction among individuals with acute traumatic brain injury and stroke. *Rehabilitation Psychology, 45*, 274–283. doi:10.1037//0090-5550.45.3.274

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review, 34*, 273–286. doi:10.1037/h0070288