Chapter 7
Step 3 of Evidence-Based Practice: Part 2—Evaluating Research Methods

Appraising the quality of research studies for practice use is Step 3 of the evidence-based practice (EBP) decision-making process. It can be a difficult task that requires professional expertise distinct from doing clinical assessment (Step 1 in EBP) or locating research resources (Step 2 in EBP). While research design (covered in Chap. 6) is one very important aspect of evaluating evidence-based medicine (EBM) and EBP research reports, it is hardly the only important methodological issue. Several other methodological choices also are important to making valid claims about treatments, diagnostic tests, or prognosis. These include the quality, diversity, size, and comprehensiveness of the sample, the validity and sensitivity of outcome measures, the definition of the treatment under investigation, and the careful use of the correct statistical tests. These methods work in tandem to yield valid and rigorous results in quantitative clinical research. In this chapter, we will review each of these issues in order.

For clarity and simplicity, we will focus on treatment outcomes in the examples used in this chapter. Readers are reminded the EBP methods can also be applied to diagnostic protocols, determination of prognoses, and even to cost-effectiveness studies. Our focus on treatment is meant to be representative, and of interest to most clinical social workers. It does not mean that the other concerns are any less important applications of EBP.

Sampling Issues

Just *who* is included in a clinical study shapes how well its results will reflect the range of persons with a particular disorder or problem. Even an experimental research design will be limited as a resource for treatment planning if it covers very few people or only people with very limited demographic characteristics. There are four key components to look for in sample selection: representativeness, diversity, size, and randomization.

J. W. Drisko, M. D. Grady, *Evidence-Based Practice in Clinical Social Work*,
Essential Clinical Social Work Series,
https://doi.org/10.1007/978-3-030-15224-6_7

Representativeness and Diversity

In quantitative research, a sample should be *representative* of the persons and setting of interest. That is, if researchers want to study a specific genetic disorder, they would ideally target all the people with the disorder. They might also include other people without the genetic disorder for comparison purposes. If there were environmental factors or cultural factors that might interact with the genetic disorder, such as diet or exposure to toxins or cultural differences, groups would be sought that vary in diet, exposure to toxins, and culture. This would represent the variety of populations impacted by the genetic disorder as best we can conceptualize them. Of course, money, time, and access might not be available to study all subgroups at once, so researchers might choose to study a smaller subset of this larger population. In this case, several studies would be necessary to obtain a sample that is fully representative of the genetic disorder and the factors we think exacerbate it or minimize it.

In clinical mental health studies, there may be a wide range of factors that could influence the effectiveness of a treatment or diagnostic test. Clinical social workers look for a wide range of biopsychosocial-spiritual factors that help understand multifaceted human problems. These include differences in gender, race, cultural or ethnicity, tribal affiliation, sexual orientation, class, age, ability, religious beliefs, legal status, genetic makeup, and geography. At times additional factors may also be important. This makes representativeness a very challenging issue for mental health researchers. Practical limitations also mean that fully representative samples may not be easy to obtain. This is especially true for minority populations and for low incidence disorders. Researchers, with input from clients and clinicians, must carefully conceptualize both their study problem formulations and the nature of their sample.

Compromises are common in sample size and representativeness due to limitations in time, funding, access, and client participation. For example, Wilt et al. (2008) report that very few RCTs on treatments for prostate cancer have been completed. They found that no type of prostate cancer treatment had then been demonstrated to be more effective than is “watchful waiting.” Side effects of the treatments were also not well identified. One reason was that men were unwilling to participate in the randomization process needed to compare different treatments. In this case, useful clinical research was limited due to the active choices of men who sought, quite understandably, to get what they believed were the best personal outcomes. Wilt et al. (2008) also note that differences in definitions and methods made the synthesis of findings across the available studies difficult to aggregate.

Where in the world samples are drawn may impact study representativeness. Glickman et al. (2009) point out that many drug trials are being “outsourced” to developing countries. They note that this raises ethical issues regarding subjecting people in these countries to the risks of research participation and allows companies to offer lower payments as incentives to participants. They also note that it is unclear if the living conditions of persons in developing countries create an appropriate sample for comparison with those in developed countries. Culture, diet, and other habits may vary. In effect, persons in developing countries *may* be an inappropriate population for sample selection in drug tests when the consumers of the drugs live in different circumstances. Other researchers argue that including a wider range of people in drug tests may benefit the clinical trial participants and others in their country as well. Selecting samples for clinical trials is complex and may raise important ethical, diversity, and representativeness issues.

Sample Size

The size of a sample is also very important. A larger sample is generally more likely to represent the population from which it was selected fully and effectively than is a smaller sample. Larger samples are also more likely to include people with diverse demographics. Larger samples tend to produce less sampling error in accurately representing an entire population. Larger samples also allow for more variety in background factors than is possible in very small samples. Further, when looking at the sample as a whole, the influence of a few "outlier" cases—those with extremely high or low scores—is also reduced as the size of the sample increases. Larger samples also generally allow for greater *statistical power*—the ability of a statistic to show an effect when it is present. Small samples may lack the power to demonstrate any effects at all. Unfortunately, there is no simple way to estimate what constitutes a "large enough" sample without considering the research question, design, sample, measures, and intended analysis type (Dattalo, 2008). The sample sizes needed to demonstrate different outcomes/effects also vary. Too small a sample may be inadequate to demonstrate the effect a researcher wants to study.

Still, it is probably a quick and useful rule that studies with less than 40 participants in total are what researchers call small samples. This would allow comparison of 20 persons in treated and control groups in an experimental research design. On the other hand, some of these small-scale studies may have samples that are quite adequate to document clinical effects, though they may not adequately include socially diverse people. Where small samples are used, having equal numbers in each group is very helpful when some statistical tests are used. Specifically having equal sized groups in an experimental comparison reduces standard error terms in these statistical analyses.

Types of Samples

Probability samples are samples in which each member of the population or sampling frame has an equal chance of being selected. A *sampling frame* is a list of potential participants used to make concrete the larger conceptual population the study seeks to address. Probability samples are intended to limit active selection bias by the researcher. Selection bias is a tendency to exclude certain cases (also called elements) from a sample. For example, persons with very severe levels of anxiety might be excluded, while persons with low or moderate levels of anxiety are included in the sample. In this example, a selection bias yields a sample which excludes persons who may make up a substantial part of the population of persons with anxiety. Such a bias is also called a *nonresponse bias*, as persons with high levels of anxiety are excluded from the study sample. Their responses will remain unknown and unstudied. The results of study, based on such a sample, will not apply well to all people who may be found in clinical practice. In other words, the result is not easily *generalized* or applied to the larger population of people with anxiety disorders. Generalization is a key goal of most experimental research.

Non-probability samples, oriented by theory or a specific research purpose, emphasize certain characteristics of sample members but do not insure equal chance of selection from the sampling frame or population. Non-probability samples may be representative, or may be very unrepresentative, of the sampling frame or population (Dattalo, 2008). When used in quantitative studies, probability samples provide a better mathematical basis for defining and limiting selection biases and nonresponsive bias than do most non-probability samples.

There are several methods for selecting probability samples. These include single or independent random sampling, systematic sampling, stratified sampling, and cluster sampling. *Simple random sampling* begins with assigning a number to each case or element in the study sampling frame. Sampling frames, however, are often not entirely inclusive of all the cases in the population of interest. A sampling frame might be the list of NASW members used as a way to define the larger, and not perfectly known, population of all social workers in the United States. The next step in simple random sampling is to use an unbiased method to select cases from the sampling frame. This is usually done using a software generated list of random numbers to select cases from the sampling frame. Selection continues, using numbers form the random listing, until the desired number of cases are selected. *Systematic random sampling* is a similar method, which begins with the random selection of a case from the sampling frame. Then every third or tenth or hundredth element is selected until the desired number of cases are selected.

Stratified sampling begins with dividing the sampling frame into groups with no shared members. For example, groups might be distinguished by ethnicity, gender differences, or age. These distinct groups are known in sampling as strata. Random sampling is then undertaken within each *stratum*. The purpose of stratified sampling is to insure adequate sampling of subgroups that are few in numbers and might not be sufficiently sampled by simple random sampling methods. Some strata may be disproportionately sampled in order to insure inclusion of enough cases from each subgroup to represent the population successfully. Finally, *cluster sampling* is used for very large populations. Cluster sampling uses existing subsets of a population to define subgroups. Random sampling is then completed on these subgroups to generate a probability sample. One common example is to use geographic areas defined by a government source to identify neighborhoods. Known, representative, neighborhoods may then be selected randomly. From within each neighborhood, cases are then selected on a random basis. Techniques to ensure probability proportionate to the size of each subgroup can be used to insure equal chance of selection for each case.

For further information on probability sampling, clinical social workers may turn to most social work research texts (Anastas, 1999; Engel & Schutt, 2016; Rubin & Babbie, 2017). However, very few social work texts offer detailed information on sampling issues related to qualitative research (Drisko, 2003). Patton (1990) and Kuzel (1999) both offer solid introductions to several varieties of qualitative sampling and their purposes.

Probability samples are used in clinical experiments or RCTs to maximize representativeness. They are often required for the appropriate use of many statistic tests. Finally, probability samples allow appropriate generalization from the sample to the larger populations from which it was drawn. Probability sampling can be a vital part of quantitative clinical research.

Increasing Statistical Power: Sample Size and Other Influences

Sample size also influences statistical analysis. Dattalo (2008) states "a study should only be conducted if it relies on a sample size that is large enough to provide an adequate and prespecified probability of finding an effect if the effect exists" (p. 16). That is, specific sample sizes are needed to generate adequate statistical power (Kraemer & Blasey, 2016). If a sample is too small, no significant effect can be demonstrated. If a sample is too large, undue and unnecessary burdens are placed on participants. The costs of completing the study also increase.

Statistical power is the probability of falsely accepting a null hypothesis when the research hypothesis is actually true (Cohen, 1988). It is defined mathematically as 1-ß, where ß (beta) is the probability of accepting the null hypothesis falsely, also called Type II error. Overall, statistical power is a function of the researcher's selected statistical significance criterion level (or *a* level) set for a specific test, the precision of measures, the type of research design, the magnitude of the effect under study, and the sample size (Dattalo, 2008; Kraemer & Blasey, 2016).

In inferential statistics, a criterion or *a* level of 0.05 (1 chance in 20) is a commonly used standard for rejecting a null hypothesis. This standard is set conservatively in order to avoid making an incorrect, "false positive," decisions, also called Type I errors. Researchers can choose to increase the criterion or *a* level to 0.10 in order to be more likely to obtain a significant result. However, in doing so the chance of false positives, known as Type I errors, is increased. Though there is no simple standard for statistical power, a value of 0.80 is widely accepted. In effect, researchers accept a 4 to 1 trade off in making a "false negative" decision, or Type II error, versus a Type I error. This means researchers are more likely to falsely accept negative results than positive (but incorrect) results. It is a careful, conservative, standard. If the magnitude of an effect is very large, a small sample might lead to a correct decision to reject the null hypothesis. But if the magnitude of the effect is small to moderate, a small sample may not be adequate to reveal it at all. Some samples are simply too small to generate adequate statistical power. This renders the result of the statistical test invalid regardless of the research question and statistical result. Increasing the sample size may be one easy and effective way to increase statistical power.

There are several methods for determining statistical power and related sample sizes (Aberson, 2010; Dattalo, 2008; Kraemer & Blasey, 2016; Murphy & Myers, 2014). These methods differ by the nature of the measures used (categorical versus continuous) and the statistics employed. There are also several computer software programs to calculate statistical power and to identify the specific sample sizes needed to obtain adequate statistical power. A power calculation should be included in all quantitative research reports.

Another way to increase statistical power is to use more structured research designs, particularly experimental and observational designs (Cook & Campbell, 1979). Such research designs reduce the number of extraneous factors that can influence the study results, thereby reducing unknown, systematic errors. Use of measures with high validly and reliability can also increase statistical power. This is because such measures reduce measurement error compared to less precise alternatives.

Overall, ideal samples for EBM/EBP research (1) will be representative of the population of interest, (2) will include human diversity in the final samples, (3) will be selected using probability sampling techniques, and (4) will be large enough in number to insure adequate statistical power. These factors are especially important in experimental or RCT research designs. Very small sample sizes (under 20 per group) warrant very careful review. This is because small samples may lack the statistical power to reveal important but modest differences in outcomes between groups. Inadequate statistical power is a greater concern when tests or measures of uncertain validity and reliability are employed. Researchers using small samples should state clearly how they determined that the sample has adequate power to produce meaningful results. This should be evident in the Methods section of the research report.

The Human Diversity Included in Study Samples

Clinicians should look carefully at the social diversity included in a study's sample. Researchers may not always report many details about the social demographics of their sample beyond age and gender. Even age, racial, and ethnic differences may be minimally detailed or omitted. Some research may unfortunately focus on populations that are not the most likely to suffer from the problem under study (Cherubini, Del Signore, Ouslander, Semla, & Michel, 2010). This may make it unclear if the sample used in an article was representative of a specific minority client whose care you are planning. It would be very helpful for clinical practice use if researchers and publishers provided greater detail about samples in research reports.

Another issue relates to *attrition* of participants as a research project continues. While study attrition is not exactly a sampling issue, it can influence the nature of the final study sample. Excellent sampling plans can be undermined when people drop out of a study, creating unequal group sizes and reducing the number of participants. Readers of research reports should look for the number and characteristics of the original, planned sample. Next, the characteristics of the researcher's final *obtained sample* should be compared to the initial "intent to treat" sample (Gupta, 2011). An intent to treat sample includes *all* participants initially randomized in a study, including those who drop out or fail to complete services. It is the most accurate and complete way to do a clinical study. Further, where follow-up measures are used, researchers should carefully document any attrition during follow-up periods. A common concern is that dropouts and clients who cannot be found for follow-up measures reduce the overall sample size and may alter the equivalence of groups compared in experimental research. Dropouts may also reduce the social diversity in a study's sample.

In mental health research, there is one final complicating issue regarding sampling. Many clients who apply for mental health services discontinue, or drop out of, services after only a few sessions. Many do not complete even planned, short-term treatments. The challenge for researchers is that it is unclear if clients who drop out have actually gotten better, gotten worse, were disappointed in the services, or left for other reasons. Knowing the reasons for dropping out could inform the research but is generally unknown and unexamined. Researchers can end up with unequal group sizes, smaller samples that undermine statistical power, and limited information of the actual effects of treatment. This can reduce the validity of experimental comparisons in mental health studies.

After a research design has been selected and the study sample defined, researchers must select tests and measures to assess key concepts. These tests and measures may define both the grouping variables that define who is treated or untreated, as well as the outcome or dependent variables that define what changes might occur.

Standardized Tests and Measures of Biopsychosocial Issues

Identifying and Locating Standardized Tests and Measures

To scientifically test if a treatment or a diagnostic protocol is effective, it is vitally important to have valid and reliable measures of the client's situation before and after the intervention. These measures may include observations, frequency counts of behaviors, spoken statements, reviews of client records, and/or standardized tests. Each data collection method has somewhat different strengths and limitations (Anastas, 1999). Standardized tests and measures are widely used in EBM/EBP research. They provide a known and replicable approach to assessing and summarizing client status and behavior.

Standardized tests and measures are developed and refined through a series of steps that helps define their validity and reliability. These characteristics are known as the *psychometric properties* of a test or measure. There are literally thousands of tests and measures that could be used in clinical social work practice. It is sometimes difficult just to understand the abbreviations used to refer to these tests and to learn their intended uses. Copyright protects most standardized tests and measures. Copyright provides protection for the intellectual property of the test creator, as well as some payment for their work. Researchers also keep some tests away from potential test takers to insure they cannot be studied and reviewed by test takers in order to influence, or even to fake, test results. One consequence of copyright protection is that the full text of test and measures may be difficult to obtain, even for practice or teaching purposes. However, some standardized tests are available in full for clinical and research uses (see Corcoran & Fischer, 2014; Hudson, 1982).

An extensive database of tests is available online, without cost, from the Educational Testing Service TestLink web site at https://www.ets.org/test_link/find_tests/. The TestLink database provides abstracts on educational tests but includes many for mental health and counseling as well. It is a fine resource to learn the basics about psychological tests. The limitation of the TestLink database is that it does not provide psychometric information to help clarify the validity and reliability of each test. Another database of tests and measures, the Health and Psychosocial Instruments (HaPI) database, is available through paid subscription or purchase only. It is found online at https://www.ebsco.com/products/research-databases/health-and-psychosocial-instruments-hapi. Many large agencies, hospitals, and social work programs have access to the HaPI database. HaPI includes links to publications about the development psychosocial tests and measures which may provide more detail than is available through TestLink. Still, neither database provides psychometric information on listed tests and measures. Neither database provides copies of tests. Both are still very useful for initially identifying tests and their intended uses.

The Buros Institute's *Mental Measurements Yearbooks* (Carlson, Geisinger, & Jonson, 2017) provide much more information about specific tests and measures. Currently in its 20th edition, these print reference books may be found in academic libraries and even in some larger public libraries. The limitation of print copies is that they may not include the latest versions of tests. (They are not exactly yearbooks; new editions appear about every 3 years.) The strength of the Buros yearbooks is that they provide details on the purposes, norming samples, range of scores, assessments validity and reliability, as well as commentary on the test. Buros Test Reviews Online allows purchase of reviews of individual tests and measures included in the print yearbooks. It is found online at http://buros.org/test-reviews-online. The test reviews online are available at http://buros.unl.edu/buros/jsp/search.jsp. Costs for purchase of individual reviews are modest.

Identifying the Specific Properties of Tests and Measures

Once you have located an appropriate test or measure, the next step is to examine its psychometric properties. These details are available in the Buros' yearbooks or online reviews, as well as in the manual available for most widely used copyrighted measures. Researchers typically provide few details about tests and measures in research reports. However, psychometric information helps readers establish the degree of confidence they should place in specific tests and measures. It also provides information about whom the test was designed to assess. This includes whether or not the test was normed on socially diverse samples and any age-related limits on use of the test. Next, we will review the attributes of tests and measures.

Sound tests and measures must be both valid and reliable. *Validity* refers to whether the measure fully captures what it is intended to measure. *Reliability* refers to whether the measure produces consistent results. Together, validity and reliability make up the key components of the *psychometric properties* of the tests and measures used in mental health research. A third factor, the *sensitivity* of a test, refers to how well it can capture the type and magnitude of changes. Sensitivity is often difficult to assess but may be very important to clinical research. Complete research reports will include the psychometric properties of all tests and measures they employ. Medical research typically focuses on nonpsychological variables using biological and physiological measures that should have strong validity and reliability.

Validity of Measures

Validity as it relates to tests in mental health research has several aspects (Campbell & Stanley, 1963). The first is *face validity* or whether or not the items (questions) that make up a test explicitly address the concepts of interest. For example, a test of marital conflict should include items that directly and overtly address different types and forms of marital conflict. A similar term is *content validity*. Content validity refers to how well the content of a test reflects the varied concepts making up a multifaceted construct. For example, measure of child maltreatment should include items about the domains of neglect, verbal abuse, sexual abuse, and physical abuse. *Construct validity* refers to the extent to which a test reflects the entire construct of interest. Some constructs may be implicit or inferred rather than directly measured in test items. For example, we would expect a test of depression to include items on mood, diminished interest in activities, sleeping patterns, feelings of worthlessness, inability to concentrate, suicidal ideation and actions, psychomotor retardation, and weight loss. These items reflect core DSM criteria for depression. A valid test must examine all of these component parts to fully cover the construct of depression as defined by DSM criteria. To exclude any one of them would reduce the construct validity of a test of depression. Note that these three aspects of validity are conceptual and require critical thinking to appraise. They also require a look at the actual items included in the measures. The absence of an important component of a construct from a measure is not (usually) captured by quantitative psychometric summaries. Clinicians need to find and look for the actual content of tests and measures to critically evaluate face, content and construct validity unless the report author includes discussion of them.

Other forms of validity are based on quantitative methods. These are collectively known as *criterion validity*. In criterion validity, the results of one test are compared to the results on another, similar, test or measure. Most texts suggest a greater than 70%, or greater than 0.70, criterion for establishing strong criterion validity. This is consistent with the way most correlation statistics are interpreted. Correlation values from 0.00 to 0.30 are generally labeled “weak” correlations, values from 0.31 to 0.70 values are labeled “moderate” correlations, and values from 0.71 to 1.00 are “strong” correlations.

In *concurrent validity* the results of similar tests are correlated with each other or to another established criterion. For example, a researcher might correlate the scores of people at similar points in time on the Beck Depression Inventory, revision II, and the Hamilton Depression Inventory. Both are measures of depression based on DSM criteria. If the results correlated highly ($r > 0.70$), the researcher could reasonably claim there was good concurrent validity between the two measures. Developers of new tests often correlate their results to the results on a more widely used test to establish the new test’s validity. *Predictive validity* refers to how well performance on a measure at one point in time predicts future performance on another measure or criterion. A researcher might find that high school grades are predictive of staying in a certain treatment program. This information might be used to screen out people with low high school grades or to examine if the program’s model and language are pitched to a higher level than is truly necessary. *Discriminant validity* refers to how well a test distinguishes between groups of different people. For example, a screening test for anxiety disorders should be able to distinguish between people likely to have an anxiety disorder from those who are unlikely to have one.

Reliability of Measures

In addition to validity, the *reliability* or consistency of a measure is vital to assessing its overall quality. Researchers and psychometricians (psychological test developers) determine the reliability of test and measures through quantitative tests. There are several methods to assess the validity of a measure. In *test-retest reliability* assessment, researchers give the same test to the same group of people at two different times, perhaps a week apart. The results of the two administrations of the test are then correlated with each other to provide a measure of test-retest reliability. Given no major environmental changes, the results are expected to correlate strongly with each other. The assumption is that the characteristics of the group will change very little in the brief time between two test administrations and that exposure to the test items will have limited impact on the results.

In *internal consistency reliability* assessment, researchers correlate the questions or items within a measure with each other. This may be done by comparing results from the first half of the test to results from the second half of the test, called *split-half reliability*. Split-half reliability assumes items are included in the test more than once and that both halves appropriately reflect the full content of interest. Other models involved complex correlations of all test items to all other items. Researchers often report internal consistency reliability using the coefficient alpha (a) statistic. Finally, *inter-rater reliability* compares the results of assessments made by two or more researchers to assess their consistency. This might include comparison of diagnoses or quantitative ratings made by clinicians. Researchers also use percentages of agreement, correlation statistics, and the Cohen's (1960) kappa statistic (*k*) to report inter-rater reliability, based on the characteristics of the test or measure.

Reporting Validity and Reliability Assessments

Due to space limitations in journal articles, many research reports provide only summary information about the psychometric properties of the measures they employ. Some include only abbreviations for tests names and cite only the test developer's manual in regard to a measure's psychometric properties. Such limited information makes it very difficult for the clinician to determine if the outcome measures used in a study are valid and reliable or truly applicable to any specific client's needs. Critical thinking is always necessary in interpreting such reports.

Clinical social workers should expect brief but detailed description of the psychometric properties of standardized tests used in EBP research. Tests should be named in full and any abbreviations used should be clearly explained. At a minimum, a citation to the test manual or other resources describing the tests purposes and psychometric properties should be clearly cited for follow-up. For example, Telch, Agras, and Linehan (2002, p. 1072) describe each standardized test they use in a single sentence followed by a full citation for further review: "Questionnaires used in this study include the Binge Eating Scale (Gormally, Black, Daston, & Rardin, 1982), a measure of severity of binge eating problems...." This is a useful start. We would argue that the validity and reliability of each test should also be described in a bit more detail to guide the reader more fully. This is often done in a very brief summary such as "the XXX depression scale has r = .81 concurrent validity when correlated to results of the widely-used YYY depression measure. The mean test-retest reliability is .76 over 4 trials with different samples." In such a summary, it is clear that the tests in use have documented validity and reliability.

Detailed information on validity and reliability is often omitted when widely used standardized tests are employed. These include tests such as the Symptom Checklist-90, the Achenbach Child Behavior Checklist, the Beck Depression Inventory, and the Hamilton Rating Scale for Depression. The drawback of this practice is that it assumes readers are familiar with the tests and measures, which is very often not that case for clinical practitioners. Further, this summary information does not specify if a standardized test has been "normed" on minority population groups, or with people who have comorbid or co-occurring disorders.

Interpreting Reports of Clinical Standardized Tests and Measures

One obvious but tricky issue in psychotherapy outcome research is to be sure the people included in a study all share the same challenge. Standardized tests are often used to verify the diagnosis of participants in research studies. For example, the

Structured Clinical Interview for DSM-III for Axis II [SCID-II] (Spitzer, Williams, Gibbon, & First, 1990) was widely used to define operationally many personality disorders. The reliability of the SCID-II was in several studies with kappa values ranging from $k = 0.02$ to 0.98 (Columbia University Biometrics Research Department, undated). The kappa values for each diagnosis included several studies with $k > 0.70$, but results were not consistent across the measures. These extremely varied results mean that across different DSM diagnoses, and evaluated using different methods, the measured reliability of the SCID-II varies widely. It may be understood as a good-enough, but far from perfect, method to determine or affirm a DSM diagnosis.

There are a wide range of tests and measures to assess client status before, during, and after treatment. For example, Binks and colleagues (2006, pp. 5–6), in their systematic review of psychological treatments for borderline personality disorder, were interested in concerns such as anxiety, depression, self-reports of self-harm, mental states, service outcomes, substance use, frequency of admission of psychiatric hospitals, or incarceration. They report these outcomes in 15 categories, including (among others) behavior, global state, mental state, substance use, economic cost, and recidivism. They go on to detail 77 specific types of outcomes, such as no change, no clinically important change, average changes, etc. (pp. 5–6). Such a wide range of variables requires a number of different techniques to assess. Some of these variables are more directly applicable to practice decision-making and immediate client needs than are others.

It is very important that measures be clearly defined and fully specified in reports. Marshall et al. (2000) found that use of poorly defined and unstandardized measures was a major limitation in their research on services for people with schizophrenia. Poorly defined outcome measures, with unknown validity and reliability, will not produce the high quality experimental results sought in EBP. While not all service outcomes can be understood in advance, it is very important that the outcome or dependent variables in an experiment be assessed using valid and reliable methods.

Some measures of status, such as length of an inpatient stay, are *direct measures* leading to frequency counts. Other measures employ scales and indices to cover a wider range of content and to get at internal states, cognition, and feelings. In all cases the process of measurement should be defined and standardized to ensure accurate assessment when used in experimental research. This enhances reader's ability to compare results across different clients and settings. Even a simple count of days of inpatient hospitalization requires a definition of just what constitutes a "day." Similarly, scales of depression or anxiety require careful construction to produce valid and reliable measurements.

Clinical rating scales come in two main types: measures of global function and disorder specific measures. For example, some studies included in Binks and colleagues' (2006) systematic review used the Global Assessment Scale [GAS] (Endicott, Spitzer, Fleiss, & Cohen, 1976) of overall psychological well-being. The GAS, completed by the clinician, rates client well-being on a 0 to 100 scale. Higher scores are positive results. The GAS is a global measure of functioning covering several domains of the patients' well-being. The Brief Psychiatric Rating Scale

[BPRS] (Overall & Gorham, 1962) was also used to assess mental state on several dimensions or subscales. Some of these 18 subscales are somatic concerns, depression, anxiety, suspiciousness, hallucinations, and grandiosity. The BPRS is scored from 18 to 126, with higher scores representing greater overall symptom severity. The BPRS, as a global standardized test, assesses both the client's stated problem and other unspoken concerns as well. Global standardized tests can help clinicians and researchers identify unstated comorbid disorders or sources of resilience and challenge that shape the client's clinical presentation.

To complement the results of global standardized tests, more narrowly focused tests are used. Tests of specific disorders or concerns are often more comprehensive in the dimensions they cover (have greater construct validity) and are often more sensitive to small differences. Thus, they are useful both to pinpoint specific client concerns and to reveal small changes that occur during treatment. The Beck Depression Inventory-II (Beck, Steer, Ball, & Ranieri, 1996) is a disorder specific standardized test that measures depression largely in terms of patient's cognitive views. Binks and colleagues (2006, p. 13) describe the BDI as measuring "supposed manifestations of depression," pointing up the importance of critical thinking and of appraising content validity! The BDI rates depression severity from 0 to 63 with higher scores indicating greater severity of depression.

The Inventory of Interpersonal Problems, Circumplex Version (Horowitz, Alden, Wiggins, & Pincus, 2000), also known as the interpersonal circumplex, measures interpersonal behavior and motives on two axes. One dimension assesses power, dominance, and need for control, while the other assesses friendliness and warmth. It is a 64-item self-report questionnaire on which each item is rated from 0 to 4 and summed up to generate an overall score. Higher scores indicate greater difficulty in interpersonal functioning. Many other disorder-specific rating scales are available for common mental health problems such as anxiety, eating disorders, and thought disorders.

Standardized tests further differ on the source of information—who fills them out—and on what information they are based. *Self-report questionnaires* are quite common. These tests are efficient and cost-effective but allow respondents to enter misleading or false information. Providing socially acceptable but inaccurate information is a widely known phenomenon. Other widely used tests are clinical rating scales based on a diagnostic interview. Such interviews must include specific content for the clinician's appraisal to be valid. Ratings made by clinicians may miss specific content that questionnaires might capture. On the other hand, clinician ratings may capture subtleties of communication and nuances missed by questionnaires. These forms of data collection are complementary.

Standardized tests also differ in sensitivity. *Test sensitivity* is the ability of a measure to correctly identify those with the concern (i.e., the true positive rate). Some standardized measures are meant more as screening tools but are also used in clinical research to measure outcomes. One example is the Achenbach Child Behavior Checklist (CBCL). The CBCL is a widely used screen test and comes in different versions for preschool (Achenbach & Rescorla, 2000) and for school-aged children (Achenbach & Rescorla, 2001). It is based on rating specific behaviors as "not true" or not evident, "sometimes true," or "always true." As a result, important changes in just one or two key behaviors may not be immediately evident in an overall CBCL score. In other words, the CBCL may lack sensitivity to small changes. Its use as an outcome measure must be carefully appraised. Optimal outcome measures have strong sensitivity to small changes. This is especially important when they are used to assess change in brief interventions.

All tests and measures used in clinical research should be reported in detail. The complete names of standardized tests should always be reported, with citations for sources. Many measures have more than one version, and multiple editions are common. At what point(s) in time the measures are completed should also be stated clearly. As noted above, the basic psychometric properties of a measure, including assessments of its validity and reliability and norming population, should be reported clearly. Limitations to the use of the measures, by age range, gender, intellectual ability, or other factors, should be clearly stated. For example, the use of adult measures with adolescents and with persons over age 65 may be invalid. Measures for children of different age ranges are also common. For progressive disorders such as Alzheimer's disease, different version of measures may be available for persons with different functional abilities. The scoring range of the measures, and whether high scores represent positive or negative results, should always be stated.

Standardized tests are increasingly available in versions useable by persons for whom English is not their first language. Bit by bit, versions of standardized tests normed for different racial and ethnic groups are being developed or identified. However, not all standardized tests have been normed on nonwhite or multicultural populations. Resources for standardized measures suitable for populations of color include Jones' (1996) and Benuto and Leany (2015) on African-American populations, Benuto (2013) on Hispanic populations, and Benuto, Thaler, and Leany (2014) on Asian populations.

For further information, most social work research texts offer good introductions to tests and measures. More detailed information on psychometrics may be found in texts by Furr and Bacharach (2007) or Rust and Golombok (2009).

Defining outcomes is a challenging process. Yet there are many test and measurement technologies available to both researchers and clinical practitioners. Still more complex is clearly defining and distinguishing among treatments and their "active" ingredients.

Defining Treatments

Standardized tests are used to assess both the baseline state (before or at the start of treatment) and later on the outcome of interventions. They are the dependent or outcome variables in EBM/EBP research. The independent variable, or the factor that leads to change in an experiment, also needs careful definition. The goal is to learn if a specific treatment causes specific changes. There are many models of biopsychosocial-spiritual interventions. Interventions also vary in modality, with individual, dyad, couple, family, group, and even community interventions available. Mental health and social service treatments also vary in complexity and in specificity. Some treatments involve several components, often delivered in a specific sequence. Other treatments may be described using a curriculum-style manual, while some are described using a set of principles but are intentionally individualized in application. Defining treatments is a very difficult undertaking. However, if the delivered treatment is not well defined, one key foundation for making cause and effect attributions is absent.

To illustrate the challenges of defining biopsychosocial therapies, we will examine Binks and colleagues' (2006, p. 4) definitions of psychological treatments for people who have borderline personality disorder (BPD). These definitions are drawn from a careful systematic review and are meant to illustrate how thoughtful researchers address the challenges of defining treatments. The authors report that they faced a "huge" number of distinct treatment types making an exhaustive listing "almost impossible" to develop (p. 4). They ended up defining six key treatment types, including cognitive-behavioral, behavioral, psychodynamic, group, miscellaneous, and standard care categories. They defined cognitive-behavioral treatments (CBT) as follows:

> A variety of interventions have been labelled CBT and it is difficult to provide a single, unambiguous definition. Recognising this, we constructed criteria we felt to be both workable and to capture the elements of good practice in CBT. In order to be classified as 'well defined' the intervention must clearly demonstrate that a component of the intervention: 1) involves the recipient establishing links between their thoughts, feelings and actions with respect to the target symptom; and 2) the correction of the person's misperceptions, irrational beliefs and reasoning biases related to the target symptom. In addition a further component of the intervention should involve either or both of the following: i) the recipient monitoring his or her own thoughts, feelings and behaviours with respect to the target symptom; and ii) the promotion of alternative ways of coping with the target symptom. All therapies that do not meet these criteria but are labelled [by the original authors as] 'CBT' or 'Cognitive Therapy' will be included as 'less well defined' CBT. (p. 4)

Here the definition of the treatment is based on a few reasonable, but broad, principles that look for the application of CBT theory in practice. Note that some CBT studies may not include enough information in their reports to be classified as CBT even if they did actually meet these standards. Note too that it would be difficult to completely replicate CBT treatments in other agency settings using this definition. Other agencies might be doing CBT according to this definition, but other factors not covered in the definition might interact to make the treatment more or less successful.

Binks and colleagues (2006, p. 4) defined psychodynamic therapy in similar fashion:

> In order to be classified as psychodynamic, the intervention must not focus on a specific presenting problem (such as aggression) but rather on the unconscious conflicts that repress the individual and need to be confronted and re-evaluated in the context of the people' [sic] adult life. The following two components had to be documented in the therapeutic intervention for the therapy to be included: a) it must explore an element of the unconscious, and b) emphasises the importance of the patient's relational interaction with the therapist.

In some measure this definition appears to define psychodynamic therapy by an absence of attention to the presenting problem, which might surprise some psychodynamically informed clinical social workers. Further, sole attention to repression seems an odd choice for treating people who have BPD as it is not a prominent defense among persons who have personality disorders. Uncovering unconscious conflicts could actually be contraindicated for people who have BPD in contemporary psychoanalytic theory and practice; supportive interventions are instead recommended (Goldstein, 1995, 2001).

The authors' intent, it seems, is to again define the therapy by how its background theory is evident in real-world practice. Yet identifying unconscious conflicts and patterns interpersonal interaction might look in practice very much like establishing links among thoughts, feelings and actions in order to change irrational (or no longer relevant) perceptions and beliefs about the target symptom. This is the same language used to define CBT!

Finally, group therapy is defined. Group therapy of course is actually a modality of treatment that can be informed by several different theories, including cognitive-behavioral and psychodynamic theories. Binks and colleagues (2006, p. 4) define group therapy as "any intervention that extends beyond the individual and specifically uses a group format in this category (e.g. family therapy and psychoanalytic group therapy). We would have included studies of therapeutic communities in this category...." Here the modality of therapy defines its key features. How specific theories are evident within the content of the group sessions is not highlighted as the defining feature for group therapy. On the other hand, theory is the defining feature used for CBT and psychodynamic therapies. Note that this definition would be quite inadequate if used to replicate any particular model of group therapy in a new setting.

To aid further clarity to the definition of treatments, researchers often report the number and duration of sessions, the qualifications of the clinicians doing the treatment, and how often supervision was provided. This information does help describe the treatments used. These descriptive efforts, too, fall short of defining treatments in a manner that allows replication in other settings. Defining mental health treatments can be very difficult.

It is interesting to note that the two therapies Binks and colleagues (2006) found to be effective in treating BPD, a psychodynamically informed partial hospital program and DBT, both included highly structured treatment programs with several components such as individual and group therapy. These shared features of the two models found to be effective were not identified in Binks and colleagues' systematic review. Instead their different theoretical foundations were emphasized. (No disrespect to Binks and colleagues is intended. We view them as going much further than do most authors in providing and explaining treatment definitions.)

Another effort to further clarify the definition of treatments or other biopsychosocial intervention processes, including diagnostic procedures, is the treatment manual. Researchers often use treatment manuals to add greater specificity to the definition of treatments.

Treatment Manuals

Treatment manuals seek to set forth the components of treatments in detail. Some go so far as to offer a curriculum, defining the tasks and activities to be completed in each session. One goal of the treatment manual is to improve the quality of treatment definitions in order to enhance the replicability and validity of clinical mental health research. Researchers view treatment manuals as an important way to increase the integrity of the intervention that causes change in experimental trails. This requires enough detail to be able to replicate the same treatment in different locations. As LeCroy (2008, p. 3) states, "treatment manuals move us closer to treatment fidelity." *Treatment fidelity* means that clinicians deliver the treatment fully as intended. It also means that different clinicians in different setting deliver the same treatment fully and consistently. This enhances replicability. Such replicability is useful in research to insure a treatment was fully delivered. In practice, it may also be promoted administratively to allow less well-trained, and less costly, providers to deliver a service. There is also no clear evidence that use of treatment manuals improves client outcomes, and there is some evidence that they do not (Truijens, Zühlke-van Hulzen, & Vanheule, 2019).

Some clinicians state that treatment manuals may undermine the individualization of therapies and other interventions to fit unique client needs, situations, and values. Ollendick, King, and Chorpita (2006) argue that treatment manuals might lead to mechanical interventions, stifling creativity and innovation. Smith (1995) called treatment manuals "cookbooks," and Silverman (1996) called them "paint by number approaches." In effect, these clinicians argue that treatment manuals omit professional expertise, a core component of EBM/EBP according to Haynes, Devereaux, and Guyatt (2002). There is a clear tension between individualizing therapy to specific and perhaps unique client needs, versus enhancing fidelity of treatment for research purposes.

In mental health, Sanderson and Woody (1995) define a treatment manual as materials that provide sufficient detail to allow a trained clinician to replicate a specific treatment. They leave unclear if description of broad psychological principles provides sufficient detail or if much greater detail is necessary. Sanderson and Woody also point out that treatment manuals are inadequate if the clinician lacks solid theoretical grounding or lacks supervised experience in the particular approach they deliver. Specifically, they point out that workshop training alone, without supervised experience, does not constitute adequate training in any therapeutic model. This view is countered, however, by manuals that claim to provide "step-by-step instructions for conducting individual and group sessions" (Center for Substance Abuse Treatment, 2007, p. 2). In such manuals, detail is substituted for professional expertise, contrary to the goals of EBM and EBP. There appear to be very different views on both the definition and optimal use of treatment manuals.

What do treatment manuals cover? Trepper et al. (n.d.) offer a treatment manual for solution-focused therapy (SFT) with individuals. Their manual details the basic tenets of SFT, how goals are set via conversations with clients, and the specific active ingredients of SFT. These ingredients include (1) a collaborative interaction between clients and clinician; (2) a positive, solution-focused stance; (3) looking for previous solutions; (4) looking for exceptions to problems; (5) using questions rather than interpretations; (6) maintaining a present time focus rather than a focus on the past; and (7) using compliments. Within each session, pre-session changes are appraised, goals are framed in terms of desired outcomes to current problems, goals are numerically scaled, and the miracle question technique may be used. The manual also includes vignettes of interactions within sessions as illustrations of the techniques.

In the SFT manual, a broad description of the therapy is combined with specification of certain techniques that make it possible to determine if the treatment was delivered in a valid and complete manner. A supervisor or a researcher could review a videotape or a transcript of a SFT session and determine if this therapy had been fully applied. Left a bit unclear is how many of these features must be present for the therapy to be called valid SFT for research purposes. For example, using many more interpretations than questions would not fit with SFT, but it is probably fine that the miracle question is not used in a specific therapy session.

Other manuals are still more detailed and prescriptive. Stark, Streusand, Krumholz, and Patel (2010) offer a manualized treatment for girls ages 9–13 and their caregivers called the ACTION program. They set forth a number of plain language themes for the program, including (1) “If you feel bad and you don’t know why, use goals skills,” (2) “If you feel bad and can change the situation, use problem solving,” and (3) “If you feel bad and it is due to negative thoughts, change the thoughts” (p. 94). Structurally, the program consists of 20 sessions of 45 to 75 minutes delivered in school to small groups of girls ($n = 2$–5). Parent training involves once a week meetings with the same therapist but for only 10 sessions. Skills emphasized in the girl’s groups include affective education, goal setting, coping skills training, and mood monitoring.

These skills are further broken down into a session-by-session format. Meeting 1 (p. 97) centers on “Introductions and discussion of pragmatics.” The objectives for meeting 1 are to: “Discus parameters of meetings. Introduce counselors and participants. Establish rationale for treatment. Discus confidentiality. Establish group rules. Build group cohesion. Establish written group incentive system.” We may assume that setting of parameters is not so unlike any other small group, but the specific rationale for the ACTION program may be. Note that building group cohesion is a universal issue for new groups but one that is very difficult to specify fully and may include some idiosyncratic components that vary from group to group.

Later meetings have different goals and progressively more focused objectives. Meeting 6 centers on “Cognition and emotion introduction to cognitive restructuring.” The objectives for meeting 6 session are to: “Demonstrate the role of cognition in emotion and behavior. Introduce connection of thoughts to feelings. Enactment of coping skills activity within session.” Over the course of the ACTION program, the group leaders teach the girl clients to be “thought detectives,” to consider if there are alternative ways to look at a problem, and to assess the evidence on which a thought is based. Several techniques fill out the objectives for Meeting 6. One such technique in the ACTION program is talking back to the “Muck Monster.” The group leaders label being unable to let go of a negative way of thinking as “being stuck in the Muck Monster.” In turn, the Muck Monster creates distance from the negative thoughts and the whole person of the client and creates a suitable opponent to challenge. The enactment within Meeting 6 is likely a direct exploration of being stuck in the Muck Monster and ways to move out of this stuck position. Such displacement of the problem and generalization of are techniques widely used across different types and theories of therapy. Later session-by-session content is also outlined and linked to related ACTION techniques. Many of the later meetings (12 to 20) include practice of the program techniques within the group setting.

It is not clear that treatment manuals fully achieve their goal of making biopsychosocial therapies more fully replicable, but they may help. Treatment manuals can make more explicit the principles and tenets, the distinguishing characteristics, and the key techniques of a treatment. This alone, however, may not allow a therapy to be fully replicated by others in a different location. Therapeutic principles and techniques overlap considerably despite differences in theory and even across treatment modalities. Individual differences in client needs, style, and comfort may require adaptations of carefully described treatment procedures. Still, treatment manuals take a useful step toward improving the validity of complex biopsychosocial interventions in order to enhance the validity of research claims made about them.

Treatment manuals are not limited to behavioral and cognitive-behavioral approaches, though they are more common for these therapies. Treatment manuals are available for certain psychodynamic psychotherapies (i.e., Clarkin, Yoemans, & Kernberg, 2006), for many behavioral and cognitive-behavioral therapies (i.e., Reilly & Shopshire, 2002; or Andrews et al., 2002), and for certain family therapies (i.e., Lock, Le Grange, Agras, & Dare, 2002). Treatment manuals for specific disorders may also include sections or chapters on different age groups or other subpopulations that are likely to be affected by the disorder (see, e.g., Benedek & Wynn, 2011 on PTSD).

The last component of appraising a research report centers on methods of analysis. For quantitative research, statistics are a vital method for decision-making. The final section in this chapter offers a review of key statistics and issues in their appropriate use.

Statistics

Statistics do not tend to be the greatest strength of many clinical social workers. While statistics are required content in most social work programs, many students do not often retain a good grasp of their use after graduation. There are many statistics, each with limiting assumptions that shape their appropriate use. We will review a number of premises for the appropriate use of statistics and point out a few key issues in interpreting statistics in research reports. It is, however, beyond this book to provide a thorough introduction or review of all statistics.

Many good introductory statistics texts are available such as Weinbach and Grinnell (2014) or Abu-Bader (2006, 2010), along with review books such as Norman and Streiner (2003).

Chapter 6 has examined how research designs shape clinical research. In interpreting research results, readers should always be clear on whether the study seeks to show differences between groups or seeks correlations among characteristics of clients. Experimental and quasi-experiment research designs explore differences between groups. Observational research designs often explore correlations among the characteristics of group members. In similar fashion, statistics fall into the same general categories: those that examine differences and those that examine correlations or associations.

Where differences are being studied, it is important that the groups being compared are as similar as possible. Comparing group differences is best achieved by using an experimental research design, but readers should further be sure the demographic characteristics (ages, genders, races, religions, etc.) and levels of functioning of the groups being compared are similar. Researchers often report comparisons of the characteristics of the groups in a clinical trial at or before the start of treatment, called a baseline. Statistics are often used to show that there is no significant difference between the treated and comparison group at baseline to document that they are similar before treatment.

Levels of Measure

Data may be either discrete or continuous. Discrete data comes only in certain finite values. If we think of "number of children," answers such as "3" or "0" make sense, but 1.5 does not. On the other hand, income is continuous data. It makes sense to have an annual income of $23,453.72, even if the cents might not matter all that much. Similarly, a scale of depression might range on a continuous scale from "0" for no depression to "20" for severely depressed. A group mean score of 12.32 for several depressed clients makes sense and allows comparison to another group with a mean score of 18.65. Most (but not all) outcome measures draw upon continuous data.

The next issue to review is the nature of the data the researchers have examined. Researchers use different statistics to examine different kinds of data. Numbers can be used to define categories with no rank order, such as "1" represents the treated group and group "2" represents the untreated control group. Measures with mutually exclusive categories and without a hierarchical ranking are called *nominal-level measures*. Numbers can also be used to establish a rough hierarchy with clear but imprecise differences among the ranks. We could use "0" to represent no formal schooling, "1" to represent some grade school, "2" to represent finished grade school, "3" to represent some middle schooling, and so forth. The higher numbers do represent more school completed, but the numbers do not reflect years of school completed in a precise and consistent manner. Measures with mutually exclusive categories and a rough hierarchy but without equal intervals between values are called *ordinal-level measures*. We can also use numbers to establish a more precise hierarchy in which the interval between the numbers represents some measured dimension. It is meaningful to distinguish between a body temperature taken by mouth of 98.6 degrees and another of 102.4 degrees. The intervals between the "tenths" of a degree are all the same and provide a scale or metric for comparison. Measures with mutually exclusive categories, a clear hierarchy of values, and equal intervals between values are called *interval-level measures*. If the scale includes a nonarbitrary zero point, we gain even more information. A body temperature of 0.00 degrees has no everyday meaning (and is not included in the range of most thermometers). But having zero dollars of annual income has a very real meaning and is much less desirable than an income of $30,000. Each dollar represents an equal and consistent increase (or decrease) in annual income. Measures with mutually exclusive categories, a clear hierarchy of values, equal intervals between values, and a nonarbitrary zero point are called *ratio-level measures*.

These differences in levels of measure are important for selecting appropriate statistics. Researchers select specific statistics in part based on the level of measure of the available data. Generally speaking, using interval- or ratio-level provides more information and allows use of more powerful statistical tests. In experiments, the independent or grouping variable must be constituted by at least nominal-level nonoverlapping categories. The dependent or outcome variable is typically interval- or ratio-level data that conveys a meaningful scale of severity. Interval- and ratio-level measures also allow for more precise scaling. While interval- and ratio-level data are more "information rich" than are nominal and ordinal-level data, any level of measures can be used as a clinical outcome (dependent) variable. For example, nominal categories (i.e., meets criteria for a DSM diagnosis or does not meet criteria) and ordinal-level data (i.e., low, moderate, or high pain severity) would both be appropriate outcome variables.

Parametric and Nonparametric Statistics: Differences in Population Distributions

Another issue that influences the selection of statistics is the nature of the distribution of values or score in the target population. All statistical tests are either parametric or nonparametric. *Parametric data* assumes that the population from which the researchers collected the sample data was a particular kind of distribution. Most often, this is to assume a normal distribution of data in the population. A normal distribution is symmetrical around the mean value, with equal "tails" on each side. Most textbooks call this the bell curve, though normal distributions can vary in look when graphed. A normal distribution means that there are roughly equal numbers of very low scores and very high scores. *Nonparametric* data distributions, on the other hand, make no assumptions about the form or parameters of a frequency distribution. In general, parametric statistics are more powerful and researchers should use them when possible. This is because nonparametric statistics are calculated using rank-order information only, which includes less specific information than do the parametric statistics.

Once the data is collected, researchers must examine the nature of the obtained sample's distribution. Data collected from a population that is assumed to be normally distributed population may prove to have different characteristics. The collected data ideally should have few "outliers" or very extreme high or low scores. In studies of small samples, a few outliers can alter the results of statistical comparisons profoundly as they increase or decrease group mean scores. In some studies, outliers are purposefully excluded from the final data analysis to avoid their strong influence on the overall results. Authors should clearly state if outliers are present and how outliers were interpreted and handled. Researchers should also review the distribution of scores in the obtained data. Distributions may be skewed or have many high or low scores, shifting them away from a symmetrical normal shape. The problem with skewed distributions is that comparing skewed and non-skewed groups may lead to results that are inaccurate. Statisticians can often transform non-normal distributions of data into a near-normal form by doing logarithmic transformations or other procedures. These transformations do not alter the relative values of scores, only the shape of their distribution. If transformations of the data distribution are undertaken, they should be clearly reported in the research report.

The Five Uses for Statistical Tests

There are five main uses for quantitative or statistical data analysis. These uses or purposes are (1) describing the characteristics of a sample or population, (2) testing for differences among groups, (3) testing for associations among variables, (4) testing for group membership, and (5) examining structure of a theory or of a measure. The first purpose is descriptive; the other four are inferential in nature.

Descriptive statistics, as the name implies, seek to (a) describe the typical or most common member of a distribution and to (b) describe the spread or dispersion found within a distribution of scores. Descriptive statistics therefore come in two types: *measures of central tendency* and *measures of dispersion*. Descriptive measures of central tendency seek to tell us about the typical member of the distribution we are studying. That is, of all the cases we have, what are the most common features and what would the typical member of the distribution look like? Descriptive measures of dispersion tell us about the variation within a distribution—how much cases differ one from the other.

Descriptive statistics are applied differentially based on the target variable's level of measure. Among descriptive measures of central tendency, only the mode can be used with a categorical or nominal measure. For an ordinal, hierarchal measure, both the mode and the median may be used. The median conveys information about both category and place in the hierarchy, so it is a bit more "information rich" than is the mode and therefore a somewhat more useful measure of central tendency. For an interval- or scaled level measure, any measure of central tendency can be used (mean, median, or mode). This is because with interval measures, we can perform mathematical operations on the data legitimately. With an interval variable, the mean is viewed as the preferable measure of central tendency because mathematical operations are used in its calculation, requiring equal intervals along the hierarchy it measures.

Measures of dispersion are all calculated using mathematical operations, so they may be used only with interval or "quantitative" measures. No measure of dispersion can be used with nominal- or ordinal-level data. Key measures are the range (maximum value minus minimum value), the variance, and the standard deviation. Skewness and kurtosis also provide information about how similar—or how different—a given distribution of scores is to a calculated "normal" distribution.

Inferential statistics, as the name implies, are used to make inferences and decisions about statistical significance. They are all based upon probability theory and compare actual, "observed" results with a mathematically constructed model that presumes no difference or no association between/among the variables under study. *Inferential statistics tells us how likely it would be to obtain a specific result if there was no difference or no association among the variables under study.* If the result is quite unlikely to have occurred by chance alone, we may say there is a statistically significant difference or correlation among the variables under study. Alas, statistics only provide probabilities and never "prove" anything absolutely. Instead they can only be said to "support" or to "fail to support" specific hypotheses about relationships among variables being studied. Still, this is a very useful technology for making decisions, especially about large groups of people.

Inferential statistics are available in many named types. Researchers select specific inferential statistics based on (a) the kind of research question being asked (about difference or association/correlation), (b) the level of measure of each variable of interest, (c) the nature of the sample (independently selected or paired/correlated selection), (d) whether the sample distribution is parametric or nonparametric, and (e) the number of variables under study. This makes it imperative to carefully think out which inferential statistic best meets your decision making needs.

Inferential statistics come in two main types: *tests of difference* and *tests of association. Tests of difference* help us decide if two or more groups differ on one or more outcome measures. Note there must be both an independent, or grouping variable (to establish the groups under comparison), *and* another dependent, or outcome, variable that reveals the extent of differences across the groups. That is, do women and men differ on average annual income? The groups are the values of gender (here limited female and male options only). The dependent variable shows difference in income. For example, the values of income establish if the groups differ through the application of statistical tests.

In *tests of association*, researchers take another approach. The goal here is to see if two variables are related, and if so, how strongly. That is, if one variable increases one value, will the other variable's value also increase or might it decrease instead? To determine if two or more variables are correlated, treatment and control groups are not needed, only values on both variables for all participants. Say a researcher measures the number of hours studied before a test and also the grades received on the test. If there is an association between the variables "hours studied" and "grades," people who studied for more hours will likely score higher than people who studied less.

Tests of association are often reversible, meaning there is no clear independent variable and no clear dependent variable. For example, the association between height and weight can be viewed from either direction. This is most common with bivariate (two variable) questions. However, with several variables under study in tests of association, we tend to think of independent variables as those that precede the dependent variable in time. For example, SAT scores precede college grades (even though they do not have much direct impact on them). Thus, we might call SAT scores the independent variable and grades the dependent variable—though the terminology gets awkward at times. It is also very important to keep in mind that even a statistically significant association does not necessarily indicate that one variable *causes* the other to change. Association or correlation does not imply cause and effect.

Multivariate statistics, based on inferential statistics, are also used to *predict group membership* and to *examine the structure of a theory* using quantitative data. *Predicting group membership* requires a large sample and interval-level data on several variables. We might want to study whether certain teens fall into "high-risk" or "low-risk" groups based on information about drug use, sexual activity, and basic mental health problems. Statistical techniques such as discriminant analysis help us predict which group one would fall based on our data.

Finally, structural equation modeling techniques, including factor analysis and principal component analysis, use interval-level data on several variables and a large sample to *explore or confirm the structure of theory and measures.* Say we wanted to create a test for depression, knowing it has several component parts such as mood problems, sleep problems, and psychomotor problems. We might collect data from people who have depression and see if these "parts" actually are elements of a general depression or if they differ enough to help identify different forms of depression (such as a predominantly sleep disturbance type which may not have much apparent mood change to it). Factor analysis takes data on each of the component parts of a theory and examines which elements (factors) maximally differ one from the other. This allows factors with similarities and distinct differences to be identified.

Choosing a Statistical Test

So, what statistical tests can researchers use and how do they select them? First, examine the nature of the research question we are asking—is it descriptive, or a question of difference or of association, or one of group membership or of theoretical structure? This is the first choice point. Next, look at the number of variables under study. Third, look at the level of measure of each variable. It may also be important to distinguish the independent and dependent variables. Fourth, review the nature of the sample. Was there independent selection versus paired or correlated selection? Fifth, for interval- or ratio-level, scaled, data, determine the nature of the data distribution. Is it a normal distribution or not? If not, can it be mathematically transformed into a near-normal distribution? From this review, researchers select a statistic that fits the mix of variables under study.

There are many charts to help researchers and statisticians pick the correct statistical test. The table that follows is adapted from Leeper's (n.d.) "Choosing the Correct Statistic." It is provided to help clinical social workers review the requirements for selecting among several widely used statistical tests. Note that the appropriate use of these tests is constrained by several factors, including the level of measure of each variable, the number of variables under study, and the nature of the distribution of the collected data (see Table 7.1).

The Misuse and Misinterpretation of Statistics in Published Reports

It should be clear by now that the correct use of statistics is a complicated process. There is a small but important literature on the misuse of statistical tests in social work and in allied mental health fields. Cowger (1984) initially described the misuse of statistical tests in the social work literature. Huxley (1986) profiled errors in the use of statistics in *The British Journal of Social Work*, Volumes 1 through 14, finding over half of the articles using statistics contained errors. Dar, Serlin, and Omer (1994) found several repeated misuses of statistical test in their review of the psychology literature between 1968 and 1988. These include inappropriate use of null hypothesis tests and p values, neglect of effect sizes, and inflation of Type I error rate through multiple comparisons. We point out these concerns to make clear to clinical social workers that statistics should not be taken simply at face value. Researchers, like all human beings, sometimes make mistakes. Critical thinking and careful attention are always required in professional endeavors.

Reporting Statistics

Statistical tests should always be reported in detail. This begins with providing enough information to allow the reader to fully determine the specific hypothesis under study. Since statistical tests actually examine the null hypothesis of no difference between groups or no association between variables, the reader should also be able to determine the null hypothesis under study. Null hypotheses are almost never stated in published reports, but they can be inferred from statements of the research

Table 7.1 Choosing a statistical test: number of independent and dependent variables, required levels of measure, and required types of data distribution

Number of dependent variables	Nature of independent variable(s)	Nature of dependent variable(s); and data distribution	Appropriate statistical test(s)
1	0 IVs (1 population)	Interval; normal	One-sample *t*-test
	"	Ordinal or interval; any distribution	One-sample median
	"	Nominal (only 2 categories); any distribution	Binomial test
	"	Nominal; any distribution	Chi-square goodness-of-fit
	1 IV with 2 levels (independent groups)	Interval; normal	2 independent sample t-test
	"	Ordinal or interval; any distribution	Wilcoxon-Mann or Whitney test
	"	Nominal; any distribution	Chi-square test
	"	Nominal; any distribution	Fisher's exact test
	1 IV with 2 or more levels (independent groups)	Interval; normal	One-way ANOVA
	"	Ordinal or interval; any distribution	Kruskal Wallis
	"	Nominal; any distribution	Chi-square test
	1 IV with 2 levels (dependent/matched groups)	Interval and normal	Paired *t*-test
	"	Ordinal or interval	Wilcoxon signed ranks test
	"	Nominal; any distribution	McNemar test
	1 IV with 2 or more levels (dependent/matched groups)	Interval and normal	One-way repeated measures ANOVA
	"	Ordinal or interval	Friedman test
	"	Nominal; any distribution	Repeated measures logistic regression
	2 or more IVs (independent groups)	Interval and normal	Factorial ANOVA
	"	Ordinal or interval	(none)
	"	Nominal; any distribution	Factorial logistic regression
	1 interval IV	Interval; normal	Correlation
	"	"	Simple linear regression

Number of dependent variables	Nature of independent variable(s)	Nature of dependent variable(s); and data distribution	Appropriate statistical test(s)
	"	Ordinal or interval; any distribution	Nonparametric correlation r_s
	"	Nominal; any distribution	Simple logistic regression
1	1 or more interval IVs and/or 1 or more nominal IVs	Interval and normal	Multiple regression
	"	"	Analysis of covariance
	"	Nominal; any distribution	Multiple logistic regression
	"	Nominal; any distribution	Discriminant analysis
2 or more	1 IV with 2 or more levels (independent groups)	Interval and normal	One-way MANOVA
2 or more	2 or more	Interval and normal	Multivariate multiple linear regression
2 sets of 2 or more	0	Interval and normal	Canonical correlation
2 or more	0	Interval and normal	Factor analysis

Adapted from "Choosing the Correct Statistic" by James Leeper of the University of Alabama College of Community and Health Sciences. (Retrieved from http://bama.ua.edu/~jleeper/627/choosestat.html)

hypothesis. This may take some effort in unpacking a complex table but should be made a bit easier by descriptions in the text as well. Note that it is almost always the case that the null hypothesis is obviously incorrect; the issue is *how unlikely* a result is to occur by chance alone. Second, after stating the hypotheses, the levels of measure for all variables should be stated if not obvious. Readers should not expect to have the level of measure for gender specified, but it should be stated for unusual tests or measures. Third, the nature of the obtained data distribution should be clearly stated. A normal distribution is required for many statistical tests. If a normal distribution is not obtained, or generated by transformation, only nonparametric statistics may be used. Fourth, the criterion level to be used to determine if results are statistically significant should be selected *before* data is collected, analyzed, and reported (Dar et al., 1994). This criterion level should be clearly stated in research reports but is often just a footnote in a table and is often mainly represented by an asterisk. This is an acceptable, if perhaps confusing, space-saving convention in publications. Readers should expect that a consistent criterion level is used throughout a study unless changes in the criterion level are explained in detail. It is inappropriate for researchers to change criterion levels without providing a rationale for such changes.

By American Psychological Association (2009) publication standards, a particular format for reporting the results of statistics is widely used. These conventions apply to both tables and text-based reports. First, the names of the variables under analysis should be clearly stated or evident in the table. Second, the name or symbol for the statistic is stated. Publishers assume that journal readers will understand the names and abbreviations for most common statistical tests. Any unusual statistical test should be explained in some detail, and a citation for more information should also be provided in the report. Third, the numerical value of the statistic is reported. Fourth, the sample size or degrees of freedom for the statistic is reported. Finally, the probability of the result is reported. It is good practice to state exact probabilities for all statistics, rather than to simply note that some are "not significant." For example, the results of an analysis of variance or F test used to compare to groups might be reported as: "A statistically significant difference on level of general anxiety was found between the treated and control groups, $F = 5.681$ (1, 85), $p = .001$." Here the value of the F statistic ($F = 5.681$) is clear, as are the degrees of freedom (1 and 85), and the precise probability value. Since probability levels vary with both the value of the statistic and the degrees of freedom (or sample size), both are reported to allow readers to verify the probability level is correct for this information.

The probability level or p value for each statistic is used to determine if the null hypothesis is to be accepted or rejected. If the p value is less than (smaller than) the criterion level in use for the study (i.e., $p = 0.003$ compared to a criterion level of $p < 0.05$), the null hypothesis is rejected. Researchers may then state that a statistically significant difference exists. Readers are often confused that reports do not directly address the null hypotheses but instead simply move on to what it implies about the research hypothesis. This too is a convention used to save space based on the assumption that professional readers should have a basic understanding of statistics.

Bear in mind that sample size influences some statistical tests. As noted above in regard to statistical power, small samples may not be able to reveal significant differences between group. On the other hand, large samples may yield significant associations even when the strength of the association is small. Readers should not confuse statistical significance with substantive or clinical significance.

To assess the magnitude of changes, effect size statistics are often reported along with tests of statistical significance. Effect size statistics complement tests of significance by more directly summarizing the size of differences between groups in experimental research (Dar et al., 1994). Effect size statistics will be examined in the next chapter.

Finally, where group differences are reported, as is common in outcome research, confidence interval should be presented along group means and probabilities. Most statistical results are presented as *point estimates* that appear quite exact. *Confidence intervals* [CI] estimate the chance that the same study, repeated with another sample taken from the same population, will yield the same results. Usually the confidence interval is established at a 95% chance that replicating the same study on a different sample will yield the same results. If the CI is narrow, the study results are more likely to be consistent when replicated. If the CI is wide, the study results are less likely to be consistent when replicated. CI ranges help the reader assess the confidence that should be placed in study results when generalizing from a single sample to the larger population. However, a confidence interval does *not* predict that the unknown, true, value of the population parameter has a defined probability of being in the confidence interval.

Summary

This chapter has examined several issues of research methodology that join with research designs to influence the validity of clinical research. Research, like clinical practice, is a complex process involving many decisions. While use of an RCT research design allows for claims of cause and effect relationships, such claims are only valid and useful if they are predicated on many other interconnected choices. The other choices include the quality and comprehensiveness of the sample, the type, validity and sensitivity of outcome measures, the quality of the definition the treatment study, and the careful use and reporting of the correct statistical tests.

Individual research reports may be integrated or synthesized to provide a summary of available research on a topic. The research designs used in individual studies may become a criterion for the inclusion or exclusion of studies from such reviews. Indeed, many summaries of research include only studies using experimental or RCT designs. Two useful resources for clinical social workers in the EBP process are meta-analysis and its elaboration into the systematic review of research studies. Examining systematic reviews will be the focus of the next chapter.

References

Aberson, C. (2010). *Applied power analysis for the behavioral sciences*. New York: Routledge Academic.

Abu-Bader, S. (2006). *Using statistical methods in social work practice*. New York: Lyceum Books.

Abu-Bader, S. (2010). *Advanced and multivariate statistical methods for social science research*. New York: Oxford University Press.

Achenbach, T., & Rescorla, L. (2000). *Manual for the ASEBA preschool forms and profiles*. Burlington, VT: University of Vermont Department of Psychiatry.

Achenbach, T., & Rescorla, L. (2001). *Manual for the ASEBA school-age forms and profiles*. Burlington, VT: University of Vermont, Research Center for Children, Youth, and Families.

American Psychological Association. (2009). *Publication manual of the American Psychological Association* (6th ed.). Washington, D.C.: American Psychological Association.

Anastas, J. W. (1999). *Research design for social work and the human services* (2nd ed.). New York: Columbia University Press.

Andrews, G., Creamer, M., Crino, R., Hunt, C., Lampe, L., & Page, A. (2002). *The treatment of anxiety disorders: Clinician guides and patient manuals*. New York: Cambridge University Press.

Beck, A., Steer, R., Ball, R., & Ranieri, W. (1996). Comparison of Beck Depression Inventories -IA and -II in psychiatric outpatients. *Journal of Personality Assessment, 67*(3), 588–597. https://doi.org/10.1207/s15327752jpa6703_13

Benedek, D., & Wynn, G. (Eds.). (2011). *Clinical manual for management of PTSD*. Arlington, VA: American Psychiatric Publishing.

Benuto, L. (2013). *Guide to psychological assessment with Hispanics*. New York: Springer Science+Business Media.

Benuto, L., & Leany, B. D. (Eds.). (2015). *Guide to psychological assessment with African Americans*. New York: Springer Science+Business Media.

Benuto, L., Thaler, N., & Leany, B. D. (Eds.). (2014). *Guide to psychological assessment with Asians*. New York: Springer Science+Business Media.

Binks, C., Fenton, M., McCarthy, L., Lee, T., Adams, C., & Duggan, C. (2006). Psychological therapies for people with borderline personality disorder. *Cochrane Database of Systematic Reviews*, (1), CD005652. https://doi.org/10.1002/14651858.CD005652

Campbell, D., & Stanley, J. (1963). *Experimental and quasi-experimental designs for research*. New York: Wadsworth.

Carlson, J., Geisinger, K., & Jonson, J. (Eds.). (2017). *The twentieth mental measurements yearbook*. Lincoln, NE: Buros Center for Testing.

Center for Substance Abuse Treatment. (2007). *Counselor's treatment manual: Matrix intensive outpatient treatment for people with stimulant use disorders.* DHHS Publication No. (SMA) 07-4152. Rockville, MD: Substance Abuse and Mental Health Services Administration. Retrieved from http://kap.samhsa.gov/products/manuals/matrix/index.htm

Cherubini, A., Del Signore, S., Ouslander, J., Semla, T., & Michel, J.-P. (2010). Fighting against age discrimination in clinical trials. *Journal of the American Geriatrics Society, 58*(9), 1791–1796.

Clarkin, J., Yoemans, F., & Kernberg, O. (2006). *Psychotherapy for borderline personality: Focusing on object relations*. Arlington, VA: American Psychiatric Publishing.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37–46.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New York: Routledge Academic.

Cook, T., & Campbell, D. (1979). *Quasi-experimentation: Design & analysis issues for field settings*. New York: Houghton Mifflin.

Corcoran, K., & Fischer, J. (2014). *Measures for clinical practice and research* (5th ed.; 2 Vols.). New York: Oxford University Press.

Cowger, C. (1984). Statistical significance tests: Scientific ritualism or scientific method? *Social Service Review, 8*(3), 358–372.

Dar, R., Serlin, R., & Omer, H. (1994). Misuse of statistical tests in three decades of psychotherapy research. *Journal of Consulting and Clinical Psychology, 62*(1), 75–82.

Dattalo, P. (2008). *Determining sample size: Balancing power, precision, and practicality*. New York: Oxford University Press.

Drisko, J. (2003, January 17). *Improving sampling strategies and terminology in qualitative research*. Paper presented at the Society for Social Work and Research Annual Meeting, Washington, D.C..

Endicott, J., Spitzer, R., Fleiss, J., & Cohen, J. (1976). The global assessment scale. A procedure for measuring overall severity of psychiatric disturbance. *Archives of General Psychiatry, 33*(6), 766–771.

Engel, R., & Schutt, R. (2016). *The practice of social work research* (4th ed.). Thousand Oaks, CA: Sage.

Furr, R. M., & Bacharach, V. (2007). *Psychometrics: An introduction*. Thousand Oaks, CA: Sage.

Glickman, S., McHutchison, J., Peterson, E., Cairns, C., Harrington, R., Califf, R., et al. (2009). Ethical and scientific implications of the globalization of clinical research. *New England Journal of Medicine, 360*, 816–823.

Goldstein, E. (1995). *Ego psychology and social work practice* (2nd ed.). New York: Free Press.

Goldstein, E. (2001). *Object relations theory and self psychology in social work practice*. New York: Free Press.

Gormally, J., Black, S., Daston, S., & Rardin, D. (1982). The assessment of binge eating severity among obese persons. *Addictive Behaviors, 7*(1), 47–55.

Gupta, S. (2011). Intention-to-treat concept: A review. *Perspectives in Clinical Research, 2*(3), 109–112.

Haynes, R., Devereaux, P., & Guyatt, G. (2002). Clinical expertise in the era of evidence-based medicine and patient choice. *Evidence-Based Medicine, 7*, 36–38.

Horowitz, L., Alden, L., Wiggins, J., & Pincus, A. (2000). *Inventory of Interpersonal Problems manual*. Odessa, FL: The Psychological Corporation.

Hudson, W. (1982). A measurement package for clinical workers. *The Journal of Applied Behavioral Science, 18*(2), 229–238.

Huxley, P. (1986). Statistical errors in papers in the *British Journal of Social Work* (Volumes 1–14). *British Journal of Social Work, 16*(6), 645–658.

Jones, R. (1996). *Handbook of test and measurements for black populations*. (2 vols.). Hampton, VA: Cobb & Henry.

Kraemer, H., & Blasey, C. (2016). *How many subjects? Statistical power analysis in research*. Thousand Oaks, CA: Sage.

Kuzel, A. (1999). Sampling in qualitative research. In B. Crabtree & W. Miller (Eds.), *Doing qualitative research* (pp. 33–46). Thousand Oaks, CA: Sage.

LeCroy, C. W. (2008). *Handbook of evidence-based treatment manuals for children and adolescents*. New York: Oxford University Press.

Leeper, J. (n.d.). *Choosing the correct statistical test*. Retrieved from http://bama.ua.edu/~jleeper/627/choosestat.html

Lock, J., Le Grange, D., Agras, W. S., & Dare, C. (2002). *Treatment manual for Anorexia Nervosa: A family-based approach*. New York: Guilford Press.

Marshall, M., Lockwood, A., Bradley, C., Adams, C., Joy, C., & Fenton, M. (2000). Unpublished rating scales: A major source of bias in randomised controlled trails of treatment for schizophrenia. *British Journal of Psychiatry, 176*, 249–252.

Murphy, K., & Myers, B. (2014). *Statistical power analysis: A simple and general model for traditional and modern hypothesis tests* (4th ed.). New York: Routledge.

Norman, G., & Streiner, D. (Eds.). (2003). *PDQ statistics* (3rd ed.). Shelton, CT: People's Medical Publishing House (PMPH).

Ollendick, T., King, N., & Chorpita, B. (2006). Empirically supported treatments for children and adolescents. In P. Kendall (Ed.), *Child and adolescent therapy: Cognitive-behavioral procedures*. New York: Guilford Press.

Overall, J., & Gorham, D. (1962). The brief psychiatric rating scale. *Psychological Reports, 10*, 799–812.

Patton, M. Q. (1990). *Qualitative evaluation and research methods* (2nd ed.). Newbury Park, CA: Sage.

Reilly, P., & Shopshire, M. (2002). *Anger management for substance abuse and mental health clients: A cognitive behavioral therapy manual*. Washington, D.C.: SAMHSA.

Rubin, A., & Babbie, E. (2017). *Research methods for social work* (8th ed.). Boston, MA: Cengage.

Rust, J., & Golombok, S. (2009). *Modern psychometrics: The science of psychological assessment* (3rd ed.). New York: Routledge.

Sanderson, W., & Woody, S. (1995). *Manuals for empirically validated treatments: A project of the Task Force on Psychological Interventions, Division of Clinical Psychology, American Psychological Association*. Washington, D.C.: American Psychological Association.

Silverman, W. (1996). Cookbooks, manuals and paint by numbers: Psychotherapy in the 1990s. *Psychotherapy, 33*, 2017–2215.

Smith, E. (1995). A passionate, rationale response to the 'manualization' of psychotherapy. *Psychological Bulletin, 22*, 36–40.

Spitzer, R., Williams, J., Gibbon, M., & First, M. (1990). *Structured Clinician Interview for DSM-III-R Axis II Disorders (SCID-II)*. Washington, D.C.: American Psychiatric Press.

Stark, K., Streusand, W., Krumholz, L., & Patel, P. (2010). Cognitive-behavioral therapy for depression: The ACTION treatment program for girls. In J. Weisz & A. Kazdin (Eds.), *Evidence-based psychotherapies for children and adolescents* (2nd ed., pp. 93–109). New York: Guilford.

Telch, C., Agras, W. S., & Linehan, M. (2002). Dialectal behavior therapy for binge eating disorder. *Journal of Consulting and Clinical Psychology, 69*(6), 1061–1065.

Trepper, T., McCollum, E., DeJong, P., Korman, H., Gingerich, W., & Franklin, C. (n.d.). *Solution focused therapy: Treatment manual for working with individuals (preliminary)*. Research Committee of the Association for Solution Focused Therapy. Retrieved from http://www.sfbta.org/Research.pdf

Truijens, F., Zühlke-van Hulzen, L., & Vanheule, S. (2019). To manualize, or not to manualize: Is that still the question? A systematic review of empirical evidence for manual superiority in psychological treatment. *Journal of Clinical Psychology, 75*, 329–343. https://doi.org/10.1002/jclp.22712

Weinbach, R., & Grinnell, R. (2014). *Statistics for social workers* (9th ed.). New York: Pearson.

Wilt, T., MacDonald, R., Rutks, I., Shamliyan, T., Taylor, B., & Kane, R. (2008). Systematic review: Comparative effectiveness and harms of treatments for clinically localized prostate cancer. *Annals of Internal Medicine, 148*(6), 435–448.