Clinical Psychology

Quantitative sophistication is increasingly central to research in clinical psychology. Both our theories and the statistical techniques available to test our hypotheses have grown in complexity over the last few decades, such that the novice clinical researcher now faces a bewildering array of analytic options. The purpose of this article is to provide a conceptual overview of the use of statistics in clinical science. The first portion of this article describes five major research questions that clinical psychology researchers commonly address and provides a brief overview of the statistical methods that frequently are employed to address each class of questions. These questions are neither exhaustive nor mutually exclusive, but rather are intended to serve as a heuristic for organizing and thinking about classes of research questions in clinical psychology and the techniques most closely associated with them. The second portion of the article articulates guiding principles that underlie the responsible use of statistics in clinical psychology.

Five Classes of Research Questions in Clinical Psychology

Defining and Measuring Constructs

Careful attention to the definition and measurement of constructs is the bread and butter of clinical research. Constructs refer to abstract psychological entities and phenomena such as depression, marital violence, genetic influences, attention to negative information, acculturation, and cognitive-behavioral therapy (CBT). We specify these unobserved variables (*see* **Latent Variable**), as well as their interrelationships, in a theoretical model (e.g., CBT might be assumed to decrease depression in one's partner, which then decreases the likelihood of marital violence in the relationship). Our measurement model (*see* **Measurement: Overview**) specifies the way in which we operationally define the constructs of interest (e.g., our 'measurement variable', or 'indicator variable', for the construct of depression might be patient scores on the Beck Depression Inventory (BDI) [4]). Finally, our analytical model refers to the way in which we statistically evaluate the hypothesized relationships between our measured variables (e.g., we might use **structural-equation modeling** (SEM), **analysis of variance** (ANOVA), or **logistic regression**). Later in this article, we discuss the importance of the consistency between these three models for making valid inferences about a theoretical model, as well as the importance of 'starting at the top' (i.e., the importance of theory for the rapid advancement of clinical research). Readers are urged to consult McFall and Townsend [36] for a more comprehensive overview of the specification and evaluation of the multiple layers of scientific models in clinical research.

Deciding how best to measure our constructs – that is, specifying the measurement model for the theoretical model of interest – is a critical first step in every clinical research project. Sometimes this step entails a challenging process of thinking logically and theoretically about how best to assess a particular construct. Consider, for example, the difficulty in defining what 'counts' as a suicide attempt. Is any dangerous personal action 'suicidal' (e.g., driving recklessly, jumping from high places, mixing barbiturates and alcohol)? Does the person have to report intending to kill herself, or are others' perceptions of her intention enough? How should intention be assessed in the very young or the developmentally delayed? Does the exhibited behavior have to be immediately life-threatening? What about life-threatening parasuicidal behaviors? Similar difficulties arise in attempting to decide how to assess physical child abuse, cognitive therapy, or an episode of overeating. These examples are intended to highlight the importance of recognizing that all phenomena of interest to clinical researchers are constructs. As a result, theoretical models of a construct and the chosen measurement models always should be distinguished – not collapsed and treated as one and the same thing – and the fit between theoretical and measurement models should be maximized.

More commonly, defining and measuring constructs entails scale development, in which researchers (a) create a set of items that are believed to assess the phenomenon or construct; (b) obtain many participants' responses to these items; and (c) use factor-analytic techniques (*see* **History of Factor Analysis: Statistical Perspective**) to reduce the

Reproduced from the *Encyclopedia of Statistics in Behavioral Science*. © John Wiley & Sons, Ltd.
ISBN: 0-470-86080-4.

complexity of the numerous items to a much smaller subset of theoretically interpretable constructs, which commonly are referred to as 'factors' or 'latent variables'. For example, Walden, Harris, and Catron [53] used factor analysis when developing 'How I Feel', a measure on which children report the frequency and intensity of five emotions (happy, sad, mad, excited, and scared), as well as how well they can control these emotions. The authors generated 30 relevant items (e.g., the extent to which children were 'scared almost all the time' during the past three months) and then asked a large number of children to respond to them. **Exploratory factor analyses** of the data indicated that three underlying factors, or constructs, accounted for much of the variability in children's responses: Positive Emotion, Negative Emotion, and Control. For example, the unobserved Negative Emotion factor accounted particularly well for variability in children's responses to the sample item above (i.e., this item showed a large factor loading on the Negative Emotion factor, and small factor loadings on the remaining two factors). One particularly useful upshot of conducting a factor analysis is that it produces **factor scores**, which index a participant's score on each of the underlying latent variables (e.g., a child who experiences chronic sadness over which she feels little control presumably would obtain a high score on the Negative Emotion factor and a lot score on the Control factor). Quantifying factor scores remains a controversial enterprise, however, and researchers who use this technique should understand the relevant issues [20]. Both Reise, Waller, and Comrey [44] and Fabrigar, Wegener, MacCallum, and Strahan [19] provide excellent overviews of the major decisions that clinical researchers must make when using exploratory factor-analytic techniques.

Increasingly, clinical researchers are making use of **confirmatory factor-analytic** techniques when defining and measuring constructs. Confirmatory approaches require researchers to specify both the number of factors and which items load on which factors *prior* to inspection and analysis of the data. Exploratory factor-analytic techniques, on the other hand, allow researchers to base these decisions in large part on what the data indicate are the best answers. Although it may seem preferable to let the data speak for themselves, the exploratory approach capitalizes on sampling variability in the data, and the resulting factor structures may be less likely to cross-validate (i.e., to hold up well in new samples of data). Thus, when your theoretical expectations are sufficiently strong to place *a priori* constraints on the analysis, it typically is preferable to use the confirmatory approach to evaluate the fit of your theoretical model to the data. Walden et al. [53] followed up the exploratory factor analysis described above by using confirmatory factor analysis to demonstrate the validity and temporal stability of the factor structure for 'How I Feel'.

Clinical researchers also use **item response theory**, often in conjunction with factor-analytic approaches, to assist in the definition and measurement of constructs [17]. A detailed description of this approach is beyond the scope of this article, but it is helpful to note that this technique highlights the importance of inspecting item-specific measurement properties, such as their difficulty level and their differential functioning as indicators of the construct of interest. For clinical examples of the application of this technique, see [27] and [30].

Cluster analysis is an approach to construct definition and measurement that is closely allied to factor analysis but exhibits one key difference. Whereas factor analysis uncovers unobserved 'factors' on the basis of the similarity of variables, cluster analysis uncovers unobserved 'typologies' on the basis of the similarity of people. Cluster analysis entails (a) selecting a set of variables that are assumed to be relevant for distinguishing members of the different typologies; (b) obtaining many participants' responses to these variables; and (c) using cluster-analytic techniques to reduce the complexity among the numerous participants to a much smaller subset of theoretically interpretable typologies, which commonly are referred to as 'clusters'. Representative recent examples of the use of this technique can be found in [21] and [24]. Increasingly, clinical researchers also are using **latent class analysis** and taxometric approaches to define typologies of clinical interest, because these methods are less descriptive and more model-based than most cluster-analytic techniques. See [40] and [6], respectively, for application of these techniques to defining and measuring clinical typologies.

Evaluating Differences between Either Experimentally Created or Naturally Occurring Groups

After establishing a valid measurement model for the particular theoretical constructs of interest, clinical

researchers frequently evaluate hypothesized group differences in dependent variables (DVs) using one of many analytical models. For this class of questions, group serves as a discrete independent or quasi-independent variable (IV or QIV). In **experimental research**, group status serves as an IV, because participants are assigned randomly to groups, as in randomized controlled trials. In **quasi-experimental research**, in contrast, group status serves as a QIV, because group differences are naturally occurring, as in psychopathology research, which examines the effect of diagnostic membership on various measures. Thus, when conducting quasi-experimental research, it often is unclear whether the QIV (a) 'causes' any of the observed group differences; (b) results from the observed group differences; or (c) has an illusory relationship with the DV (e.g., a third variable has produced the correlation between the QIV and the DV). Campbell and Stanley [9] provide an excellent overview of the theoretical and methodological issues surrounding the distinction between quasi-experimental and experimental research and describe the limits of causality inferences imposed by the use of quasi-experimental research designs.

In contrast to the IV or QIV, the DVs can be continuous or discrete and are presumed to reflect the influence of the IV or QIV. Thus, we might be interested in (a) evaluating differences in perfectionism (the DV) for patients who are diagnosed with anorexia versus bulimia (a QIV, because patients are not assigned randomly to disorder type); (b) examining whether the frequency of rehospitalization (never, once, two or more times) over a two-year period (the DV) varies for patients whose psychosis was or was not treated with effective antipsychotic medication during the initial hospitalization (an IV, if drug assignment is random); (c) investigating whether the rate of reduction in hyperactivity (the DV) over the course of psychopharmacological treatment with stimulants is greater for children whose parents are assigned randomly to implement behavioral-modification programs in their homes (an IV); (d) assessing whether the time to a second suicide attempt (the DV) is shorter for patients who exhibit marked, rather than minimal, impulsivity (a QIV); or (e) evaluating whether a 10-day behavioral intervention versus no intervention (an IV) reduces the overall level of a single child's disruptive behavior (the DV).

What sets apart this class of questions about the influence of an IV or QIV on a DV is the discreteness of the predictor; the DVs can be practically any statistic, whether means, proportions, frequencies, slopes, correlations, time until a particular event occurs, and so on. Thus, many statistical techniques aim to address the same meta-level research question about group differences but they make different assumptions about the nature of the DV. For example, clinical researchers commonly use ANOVA techniques to examine group differences in means (perhaps to answer question 1 above); chi-square or **log-linear** approaches to evaluate group differences in frequencies (question 2; see [52]); **growth-curve** or multilevel modeling (MLM) (*see* **Hierarchical Models**) techniques to assess group differences in the intercept, slope, or acceleration parameters of a regression line (question 3; see [48] for an example); survival analysis to investigate group differences in the time to event occurrence, or 'survival time' (question 4; see [7] and [8]); and **interrupted time-series analysis** to evaluate the effect of an intervention on the level or slope of a single participant's behavior within a multiple-baseline design (question 5; see [42] for an excellent example of the application of this approach). Thus, these five very different analytical models all aim to evaluate very similar theoretical models about group differences. A common extension of these analytical models provides simultaneous analysis of two or more DVs (e.g., **Multivariate Analysis of Variance (MANOVA)** evaluates mean group differences in two or more DVs).

Many analyses of group differences necessitate inclusion of one or more covariates, or variables other than the IV or QIV that also are assumed to influence the DV and may correlate with the predictor. For example, a researcher might be interested in evaluating the influence of medication compliance (a QIV) on symptoms (the DV), apart from the influence of social support (the covariate). In this circumstance, researchers commonly use **Analysis of Covariance** (ANCOVA) to 'control for' the influence of the covariate on the DV. If participants are assigned randomly to levels of the IV, then ANCOVA can be useful for increasing the **power** of the evaluation of the effect of the IV on the DV (i.e., a true effect is more likely to be detected). If, however, participants are not assigned randomly to IV levels and the groups differ on the covariate – a common circumstance in clinical research and a likely

characteristic of the example above – then ANCOVA rarely is appropriate (i.e., this analytical model likely provides an invalid assessment of the researcher's theoretical model). This is an underappreciated matter of serious concern in psychopathology research, and readers are urged to consult [39] for an excellent overview of the relevant substantive issues.

Predicting Group Membership

Clinical researchers are interested not only in examining the effect of group differences on variables of interest (as detailed in the previous section) but also in predicting group differences. In this third class of research questions, group differences become the DV, rather than the IV or QIV. We might be interested in predicting membership in diagnostic categories (e.g., schizophrenic or not) or in predicting important discrete clinical outcomes (e.g., whether a person commits suicide, drops out of treatment, exhibits partner violence, reoffends sexually after mandated treatment, or holds down a job while receiving intensive case-management services). In both cases, the predictors might be continuous, discrete, or a mix of both. **Discriminant function analysis** (DFA) and **logistic regression** techniques commonly are used to answer these kinds of questions. Note that researchers use these methods for a purpose different than that of researchers who use the typology-definition methods discussed in the first section (e.g., cluster analysis, latent class analysis); the focus in this section is on the *prediction* of group membership (which already is known before the analysis), rather than the *discovery* of group membership (which is unknown at the beginning of the analysis).

DFA uses one or more weighted linear combinations of the predictor variables to predict group membership. For example, Hinshaw, Carte, Sami, Treuting, and Zupan [22] used DFA to evaluate how well a class of 10 neuropsychiatric variables could predict the presence or absence of attention-deficit/hyperactivity disorder (ADHD) among adolescent girls. Prior to conducting the DFA, Hinshaw and colleagues took the common first step of using MANOVA to examine whether the groups differed on a linear combination of the class of 10 variables (i.e., they first asked the group-differences question that was addressed in the previous section). Having determined that the groups differed on the class of variables, as well as on each of the 10 variables in isolation, the authors then used DFA to predict whether each girl did or did not have ADHD. DFA estimated a score for each girl on the weighted linear combination (or discriminant function) of the predictor variables, and the girl's predicted classification was based on whether her score cleared a particular cutoff value that also was estimated in the analysis. The resulting discriminant function, or prediction equation, then could be used in other samples or studies to predict the diagnosis of girls for whom ADHD status was unknown. DFA produces a two-by-two classification table, in which the two dimensions of the table are 'true' and 'predicted' states (e.g., the presence or absence of ADHD). Clinical researchers use the information in this table to summarize the predictive power of the collection of variables, commonly using a percent-correct index, a combination of sensitivity and specificity indices, or a combination of positive and negative predictive power indices. The values of these indices frequently vary as a function of the relative frequency of the two states of interest, as well as the cutoff value used for classification purposes, however. Thus, researchers increasingly are turning to alternative indices without these limitations, such as those drawn from **signal-detection theory** [37].

Logistic regression also examines the prediction of group membership from a class of predictor variables but relaxes a number of the restrictive assumptions that are necessary for the valid use of DFA (e.g., multivariate normality, linearity of relationships between predictors and DV, and homogeneity of variances within each group). Whereas DFA estimates a score for each case on a weighted linear combination of the predictors, logistic regression estimates the probability of one of the outcomes for each case on the basis of a nonlinear (logistic) transformation of a weighted linear combination of the predictors. The predicted classification for a case is based on whether the estimated probability clears an estimated cutoff. Danielson, Youngstrom, Findling, and Calabrese [16] used logistic regression in conjunction with signal-detection theory techniques to quantify how well a behavior inventory discriminated between various diagnostic groups. At this time, logistic regression techniques are preferred over DFA methods, given their less-restrictive assumptions.

Evaluating Theoretical Models That Specify a Network of Interrelated Constructs

As researchers' theoretical models for a particular clinical phenomenon become increasingly sophisticated and complex, the corresponding analytical models also increase in complexity (e.g., evaluating a researcher's theoretical models might require the simultaneous estimation of multiple equations that specify the relationships between a network of variables). At this point, researchers often turn to either multiple-regression models (MRM) (*see* **Multiple Linear Regression**) or SEM to formalize their analytical models. In these models, constructs with a single measured indicator are referred to as measured (or manifest) variables; this representation of a construct makes the strong assumption that the measured variable is a perfect, error-free indicator of the underlying construct. In contrast, constructs with multiple measured indicators are referred to as latent variables; the assumption in this case is that each measured variable is an imperfect indicator of the underlying construct and the inclusion of multiple indicators helps to reduce error.

MRM is a special case of SEM in which all constructs are treated as measured variables and includes single-equation multiple-regression approaches, **path-analytic methods**, and **linear multilevel models** techniques. Suppose, for example, that you wanted to test the hypothesis that the frequency of negative life events influences the severity of depression, which in turn influences physical health status. MRM would be sufficient to evaluate this theoretical model if the measurement model for each of these three constructs included only a single variable. SEM likely would become necessary if your measurement model for even one of the three constructs included more than one measured variable (e.g., if you chose to measure physical health status with scores on self-report scale as well as by medical record review, because you thought that neither measure in isolation reliably and validly captured the theoretical construct of interest). Estimating SEMs requires the use of specialized software, such as LISREL, AMOS, M-PLUS, Mx, or EQS (*see* **Structural Equation Modeling: Software**).

Two types of multivariate models that are particularly central to the evaluation and advancement of theory in clinical science are those that specify either **mediation** or **moderation** relationships between three or more variables [3]. Mediation hypotheses specify a mechanism (B) through which one variable (A) influences another (C). Thus, the example in the previous paragraph proposes that severity of depression (B) mediates the relationship between the frequency of negative life events (A) and physical health (C); in other words, the magnitude of the association between negative life events and physical health should be greatly reduced once depression enters the mix. The strong version of the mediation model states that the A-B-C path is causal and complete – in our example, that negative life events cause depression, which in turn causes a deterioration in physical health – and that the relationship between A and C is completely accounted for by the action of the mediator. Complete mediation is rare in social science research, however. Instead, the weaker version of the mediation model is typically more plausible, in which the association between A and C is reduced significantly (but not eliminated) once the mediator is introduced to the model.

In contrast, moderation hypotheses propose that the magnitude of the influence of one variable (A) on another variable (C) depends on the value of a third variable (B) (i.e., moderation hypotheses specify an interaction between A and B on C). For example, we might investigate whether socioeconomic status (SES) (B) moderates the relationship between negative life events (A) and physical health (C). Conceptually, finding a significant moderating relationship indicates that the A−C relationship holds only for certain subgroups in the population, at least when the moderator is discrete. Such subgroup findings are useful in defining the boundaries of theoretical models and guiding the search for alternative theoretical models in different segments of the population.

Although clinical researchers commonly specify mediation and moderation theoretical models, they rarely design their studies in such a way as to be able to draw strong inferences about the hypothesized theoretical models (e.g., many purported mediation models are evaluated for data collected in **cross-sectional designs** [54], which raises serious concerns from both a logical and data-analytic perspective [14]). Moreover, researchers rarely take all the steps necessary to evaluate the corresponding analytical models. Greater attention to the relevant literature on appropriate statistical evaluation of mediation and moderation

hypotheses should enhance the validity of our inferences about the corresponding theoretical models [3, 23, 28, 29].

In addition to specifying mediating or moderating relationships, clinical researchers are interested in networks of variables that are organized in a nested or hierarchical fashion. Two of the most common **hierarchical**, or multilevel, data structures are (a) nesting of individuals within social groups or organizations (e.g., youths nested within classrooms) or (b) nesting of observations within individuals (e.g., multiple symptoms scores over time nested within patients). Prior to the 1990s, options for analyzing these nested data structures were limited. Clinical researchers frequently collapsed multilevel data into a flat structure (e.g., by disaggregating classroom data to the level of the child or by using difference scores to measure change within individuals). This strategy resulted in the loss of valuable information contained within the nested data structure and, in some cases, violated assumptions of the analytic methods (e.g., if multiple youths are drawn from the same classroom, their scores will likely be correlated and violate independence assumptions). In the 1990s, however, advances in statistical theory and computer power led to the development of MLM techniques. Conceptually, MLM can be thought of as hierarchical multiple regression, in which regression equations are estimated for the smallest (or most nested) unit of analysis and then the parameters of these regression equations are used in second-order analyses. For example, a researcher might be interested in both individual-specific and peer-group influences on youth aggression. In an MLM analysis, two levels of regression equations would be specified: (a) a first-level equation would specify the relationship of individual-level variables to youth aggression (e.g, gender, attention problems, prior history of aggression in a different setting, etc.); and (b) a second-level equation would predict variation in these individual regression parameters as a function of peer-group variables (e.g., the effect of average peer socioeconomic status (SES) on the relationship between gender and aggression). In practice, these two levels are estimated simultaneously. However, given the complexity of the models that can be evaluated using MLM techniques, it is frequently useful to map out each level of the MLM model separately. For a through overview of MLM techniques and available statistical packages, see the recent text by Raudenbush and Byrk [43], and for recent applications of MLM techniques in the clinical literature, see [41] and [18].

Researchers should be forewarned that numerous theoretical, methodological, and statistical complexities arise when specifying, estimating, and evaluating an analytical model to evaluate a hypothesized network of interrelated constructs, particularly when using SEM methods. Space constraints preclude description of these topics, but researchers who wish to test more complex theoretical models are urged to familiarize themselves with the following particularly important issues: (a) Evaluation and treatment of missing-data patterns; (b) assessment of power for both the overall model and for individual parameters of particular interest; (c) the role of capitalization on chance and the value of cross-validation when respecifying poorly fitting models; (d) the importance of considering different models for the network of variables that make predictions identical to those of the proposed theoretical model; (e) the selection and interpretation of appropriate fit indices; and (f) model-comparison and model-selection procedures (e.g., [2, 14, 25, 32, 33, 34, 51]). Finally, researchers are urged to keep in mind the basic maxim that the strength of causal inferences is affected strongly by research design, and the experimental method applied well is our best strategy for drawing such inferences. MRM and SEM analytical techniques often are referred to as causal models, but we deliberately avoid that language here. These techniques may be used to analyze data from a variety of experimental or quasi-experimental research designs, which may or may not allow you to draw strong causal inferences.

Synthesizing and Evaluating Findings Across Studies or Data Sets

The final class of research questions that we consider is research synthesis or **meta-analysis**. In meta-analyses, researchers describe and analyze empirical findings across studies or datasets. As in any other research enterprise, conducting a meta-analysis (a) begins with a research question and statement of hypotheses; (b) proceeds to data collection, coding, and transformation; and (c) concludes with analysis and interpretation of findings. Meta-analytic investigations differ from other studies in that the unit of data collection is the *study* rather than the participant. Accordingly, 'data collection' in meta-analysis is typically an exhaustive, well-documented literature

search, with predetermined criteria for study inclusion and exclusion (e.g., requiring a minimum sample size or the use of random assignment). Following initial data collection, researchers develop a coding scheme to capture the critical substantive and methodological characteristics of each study, establish the reliability of the system, and code the findings from each investigation. The empirical results of each investigation are transformed into a common metric of effect sizes (see [5] for issues about such transformations). Effect sizes then form the unit of analysis for subsequent statistical tests. These statistical analyses may range from a simple estimate of a population effect size in a set of homogenous studies to a complex multivariate model designed to explain variability in effect sizes across a large, diverse literature.

Meta-analytic inquiry has become a substantial research enterprise within clinical psychology, and results of meta-analyses have fueled some of the most active debates in the field. For example, in the 1980s and 1990s, Weisz and colleagues conducted several reviews of the youth therapy treatment literature, estimated population effect sizes for the efficacy of treatment versus control conditions, and sought to explain variability in these effect sizes in this large and diverse treatment literature (e.g., [56]). Studies included in the meta-analyses were coded for theoretically meaningful variables such as treatment type, target problem, and youth characteristics. In addition, studies were classified comprehensively in terms of their methodological characteristics – from the level of the study (e.g., sample size, type of control group) down to the level of each individual outcome measure, within each treatment group, within each study (e.g., whether a measure was an unnecessarily reactive index of the target problem). This comprehensive coding system allowed the investigators to test the effects of the theoretical variables of primary interest as well as to examine the influence of methodological quality on their findings. Results of these meta-analyses indicated that (a) structured, behavioral treatments outperformed unstructured, nonbehavioral therapies across the child therapy literature; and (b) psychotherapy in everyday community clinic settings was more likely to entail use of nonbehavioral treatments and to have lower effect sizes than those seen in research studies of behavioral therapies (e.g., [55]). The debate provoked by these meta-analytic findings continues, and the results have spurred research on the moderators of therapy effects and the dissemination of evidence-based therapy protocols to community settings.

As our example demonstrates, meta-analysis can be a powerful technique to describe and explain variability in findings across an entire field of inquiry. However, meta-analysis is subject to the same limitations as other analytic techniques. For example, the effects of a meta-analysis can be skewed by biased sampling (e.g., an inadequate literature review), use of a poor measurement model (e.g., an unreliable scheme for coding study characteristics), low power (e.g., an insufficiently large literature to support testing cross-study hypotheses), and data-quality problems (e.g., a substantial portion of the original studies omit data necessary to evaluate meta-analytic hypotheses, such as a description of the ethnicity of the study sample). Furthermore, most published meta-analyses do not explicitly model the nested nature of their data (e.g., effect sizes on multiple symptom measures are nested within treatment groups, which are nested within studies). Readers are referred to the excellent handbook by Cooper and Hedges [15] for a discussion of these and other key issues involved in conducting a meta-analysis and interpreting meta-analytic data.

Overarching Principles That Underlie the Use of Statistics in Clinical Psychology

Having provided an overview of the major research questions and associated analytical techniques in clinical psychology, we turn to a brief explication of four principles and associated corollaries that characterize the responsible use of statistics in clinical psychology. The intellectual history of these principles draws heavily from the work and insight of such luminaries as **Jacob Cohen**, Alan Kazdin, Robert McCallum, and Paul Meehl. Throughout this section, we refer readers to more lengthy articles and texts that expound on these principles.

Principle 1: The specification and evaluation of theoretical models is critical to the rapid advancement of clinical research.

Corollary 1: Take specification of theoretical, measurement, and analytical models seriously. As theoretical models specify unobserved constructs and their interrelationships (see earlier section on defining

and measuring constructs), clinical researchers must draw inferences about the validity of their theoretical models from the fit of their analytical models. Thus, the strength of researchers' theoretical inferences depends critically on the consistency of the measurement and analytical models with the theoretical models [38]. Tightening the fit between these three models may preclude the use of 'off-the-shelf' measures or analyses, when existing methods do not adequately capture the constructs or their hypothesized interrelationships. For example, although more than 25 years of research document the outstanding psychometric properties of the BDI, the BDI emphasizes the cognitive and affective aspects of the construct of depression more than the vegetative and behavioral aspects. This measurement model may be more than sufficient for many investigations, but it would not work well for others (e.g., a study targeting sleep disturbance). Neither measurement nor analytical models are 'assumption-free', so we must attend to the psychometrics of measures (e.g., their reliability and validity), as well as to the assumptions of analytical models. Additionally, we must be careful to maintain the distinctions among the three models. For example, clinical researchers tend to collapse the theoretical and measurement models as work progresses in a particular area (e.g., we reify the construct of depression as the score on the BDI). McFall and Townsend [36] provide an eloquent statement of this and related issues.

Corollary 2: Pursue theory-driven, deductive approaches to addressing research questions whenever possible, and know the limitations of relying on more inductive strategies. Ad hoc storytelling about the results of innumerable exploratory data analyses is a rampant research strategy in clinical psychology. Exploratory research and data analysis often facilitate the generation of novel theoretical perspectives, but it is critical to replicate the findings and examine the validity of a new theoretical model further before taking it too seriously.

Principle 2: The heart of the clinical research enterprise lies in model (re-)specification, evaluation, and comparison.

Corollary 1: Identify the best model from a set of plausible alternatives, rather than evaluating the adequacy of a single model. Clinical researchers often evaluate a hypothesized model only by comparing it to models of little intrinsic interest, such as a null model that assumes that there is no relationship between the variables or a saturated model that accounts perfectly for the observed data. Serious concerns still may arise in regard to a model that fits significantly better than the null model and nonsignificantly worse than the saturated model, however, (see [51] for an excellent overview of the issues that this model-fitting strategy raises). For example, a number of equivalent models may exist that make predictions identical to those of the model of interest [34]. Alternatively, nonequivalent alternative models may account as well or better for the observed data. Thus, methodologists now routinely recommend that researchers specify and contrast competing theoretical models (both equivalent and nonequivalent) because this forces the researcher to specify and evaluate a variety of theoretically based explanations for the anticipated findings [34, 51].

Corollary 2: Model modifications may increase the validity of researchers' theoretical inferences, but they also may capitalize on sampling variability. When the fit of a model is less than ideal, clinical researchers often make post hoc modifications to the model that improve its fit to the observed data set. For example, clinical researchers who use SEM techniques often delete predictor variables, modify the links between variables, or alter the pattern of relationships between error terms. Other analytic techniques also frequently suffer from similar overfitting problems (e.g., stepwise regression (*see* **Regression Models**), DFA). These data-driven modifications improve the fit of the model significantly and frequently can be cast as theoretically motivated. However, these changes may do little more than capitalize on systematic but idiosyncratic aspects of the sample data, in which case the new model may not generalize well to the population as a whole [33, 51]. Thus, it is critical to cross-validate respecified models by evaluating their adequacy with data from a new sample; alternatively, researchers might develop a model on a randomly selected subset of the sample and then cross-validate the resulting model on the remaining participants. Moreover, to be more certain that the theoretical assumptions about the need for the modifications are on target, it is important to evaluate the novel theoretical implications of the modified model with additional data sets.

Principle 3: Mastery of research design and the mechanics of statistical techniques is critical to the validity of researchers' statistical inferences.

Corollary 1: Know your data. Screening data is a critical first step in the evaluation of any analytical model. Inspect and address patterns of missing data (e.g., pair-wise deletion, list-wise deletion, estimation of missing data). Evaluate the assumptions of statistical techniques (e.g., normality of distributions of errors, absence of outliers, linearity, homogeneity of variances) and resolve any problems (e.g., make appropriate data transformations, select alternative statistical approaches). Tabachnick and Fidell [50] provide an outstanding overview of the screening process in the fourth chapter of their multivariate text.

Corollary 2: Know the power of your tests. Jacob Cohen [10] demonstrated more than four decades ago that the **power** to detect hypothesized effects was dangerously low in clinical research, and more recent evaluations have come to shockingly similar conclusions [47, 49]. Every clinical researcher should understand how sample size, effect size, and α affect power; how low power increases the likelihood of erroneously rejecting our theoretical models; and how exceedingly high power may lead us to retain uninteresting theoretical models. Cohen's [12] power primer is an excellent starting place for the faint of heart.

Corollary Three: Statistics can never take you beyond your methods. First, remember GIGO (garbage in – garbage out): Running statistical analyses on garbage measures invariably produces garbage results. Know and care deeply about the psychometric properties of your measures (e.g., various forms of reliability, validity, and **generalizability**; see [26] for a comprehensive overview). Second, note that statistical techniques rarely can eliminate confounds in your research design (e.g., it is extremely difficult to draw compelling causal inferences from quasi-experimental research designs). If your research questions demand quasi-experimental methods, familiarize yourself with designs that minimize threats to the internal and external validity of your conclusions [9, 26].

Principle 4: Know the limitations of Null-Hypothesis Statistical Testing (NHST).

Corollary 1: The alternative or research hypotheses tested within the NHST framework are very imprecise and almost always true at a population level. With enough power, almost any two means will differ significantly, and almost any two variables will show a statistically significant correlation. This weak approach to the specification and evaluation of theoretical models makes it very difficult to reject or falsify a theoretical model, or to distinguish between two theoretical explanations for the same phenomena. Thus, clinical researchers should strive to develop and evaluate more precise and risky predictions about clinical phenomena than those traditionally examined with the NHST framework [11, 13, 31, 38]. When the theoretical models in a particular research area are not advanced enough to allow more precise predictions, researchers are encouraged to supplement NHST results by presenting **confidence intervals** around sample statistics [31, 35].

Corollary 2: P *values do not tell you the likelihood that either the null or alternative hypothesis is true*. *P* values specify the likelihood of observing your findings if the null hypothesis is true – *not* the likelihood that the null hypothesis is true, given your findings. Similarly, (1.0 – p) is not equivalent to the likelihood that the alternative hypothesis is true, and larger values of (1.0 – p) do not mean that the alternative hypothesis is more likely to be true [11, 13]. Thus, as Abelson [1] says, 'Statistical techniques are aids to (hopefully wise) judgment, not two-valued logical declarations of truth or falsity' (p. 9–10).

Corollary 3: Evaluate practical significance as well as statistical significance. The number of 'tabular asterisks' in your output (i.e., the level of significance of your findings) is influenced strongly by your sample size and indicates more about reliability than about the practical importance of your findings [11, 13, 38]. Thus, clinical researchers should report information on the practical significance, or magnitude, of their effects, typically by presenting effect-size indices and the confidence intervals around them [13, 45, 46]. Researchers also should evaluate the adequacy of an effect's magnitude by considering the

domain of application (e.g., a small but reliable effect size on mortality indices is nothing to scoff at!).

Conclusions

Rapid advancement in the understanding of complex clinical phenomena places heavy demands on clinical researchers for thoughtful articulation of theoretical models, methodological expertise, and statistical rigor. Thus, the next generation of clinical psychologists likely will be recognizable in part by their quantitative sophistication. In this article, we have provided an overview of the use of statistics in clinical psychology that we hope will be particularly helpful for students and early career researchers engaged in advanced statistical and methodological training. To facilitate use for teaching and training purposes, we organized the descriptive portion of the article around core research questions addressed in clinical psychology, rather than adopting alternate organizational schemes (e.g., grouping statistical techniques on the basis of mathematical similarity). In the second portion of the article, we synthesized the collective wisdom of statisticians and methodologists who have been critical in shaping our own use of statistics in clinical psychological research. Readers are urged to consult the source papers of this section for thoughtful commentary relevant to all of the issues raised in this article.

References

[1] Abelson, R.P. (1995). *Statistics as Principled Argument*, Lawrence Erlbaum Associates, Hillsdale.

[2] Allison, P.D. (2003). Missing data techniques for structural equation modeling, *Journal of Abnormal Psychology* **112**, 545–557.

[3] Baron, R.M. & Kenny, D.A. (1986). The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations, *Journal of Personality and Social Psychology* **51**, 1173–1182.

[4] Beck, A.T., Steer, R.A. & Brown, G.K. (1996). *Manual for the Beck Depression Inventory*, 2nd Edition, The Psychological Corporation, San Antonio.

[5] Becker, B.J., ed. (2003). Special section: metric in meta-analysis, *Psychological Methods* **8**, 403–467.

[6] Blanchard, J.J., Gangestad, S.W., Brown, S.A. & Horan, W.P. (2000). Hedonic capacity and schizotypy revisited: a taxometric analysis of social anhedonia, *Journal of Abnormal Psychology* **109**, 87–95.

[7] Brent, D.A., Holder, D., Kolko, D., Birmaher, B., Baugher, M., Roth, C., Iyengar, S. & Johnson, B.A. (1997). A clinical psychotherapy trial for adolescent depression comparing cognitive, family, and supportive therapy, *Archives of General Psychiatry* **54**, 877–885.

[8] Brown, G.K., Beck, A.T., Steer, R.A. & Grisham, J.R. (2000). Risk factors for suicide in psychiatric outpatients: a 20-year prospective study, *Journal of Consulting & Clinical Psychology* **68**, 371–377.

[9] Campbell, D.T. & Stanley, J.C. (1966). *Experimental and Quasi-experimental Designs for Research*, Rand McNally, Chicago.

[10] Cohen, J. (1962). The statistical power of abnormal-social psychological research: a review, *Journal of Abnormal and Social Psychology* **65**, 145–153.

[11] Cohen, J. (1990). Things I have learned (so far), *American Psychologist* **45**, 1304–1312.

[12] Cohen, J. (1992). A power primer, *Psychological Bulletin* **112**, 155–159.

[13] Cohen, J. (1994). The earth is round, *American Psychologist* **49**, 997–1003.

[14] Cole, D.A. & Maxwell, S.E. (2003). Testing mediational models with longitudinal data: questions and tips in the use of structural equation modeling, *Journal of Abnormal Psychology* **112**, 558–577.

[15] Cooper, H. & Hedges, L.V., eds (1994). *The Handbook of Research Synthesis*, Sage, New York.

[16] Danielson, C.K., Youngstrom, E.A., Findling, R.L. & Calabrese, J.R. (2003). Discriminative validity of the general behavior inventory using youth report, *Journal of Abnormal Child Psychology* **31**, 29–39.

[17] Embretson, S.E. & Reise, S.P. (2000). *Item Response Theory for Psychologists*, Lawrence Erlbaum Associates, Hillsdale.

[18] Espelage, D.L., Holt, M.K. & Henkel, R.R. (2003). Examination of peer-group contextual effects on aggression during early adolescence, *Child Development* **74**, 205–220.

[19] Fabrigar, L.R., Wegener, D.T., MacCallum, R.C. & Strahan, E.J. (1999). Evaluating the use of exploratory factor analysis in psychological research, *Psychological Methods* **4**, 272–299.

[20] Grice, J.W. (2001). Computing and evaluating factor scores, *Psychological Methods* **6**, 430–450.

[21] Grilo, C.M., Masheb, R.M. & Wilson, G.T. (2001). Subtyping binge eating disorder, *Journal of Consulting and Clinical Psychology* **69**, 1066–1072.

[22] Hinshaw, S.P., Carte, E.T., Sami, N., Treuting, J.J. & Zupan, B.A. (2002). Preadolescent girls with attention-deficit/hyperactivity disorder: II. Neuropsychological performance in relation to subtypes and individual classification, *Journal of Consulting and Clinical Psychology* **70**, 1099–1111.

[23] Holmbeck, G.N. (1997). Toward terminological, conceptual, and statistical clarity in the study of mediators and

moderators: examples from the child-clinical and pediatric psychology literature, *Journal of Consulting and Clinical Psychology* **65**, 599–610.

[24] Holtzworth-Munroe, A., Meehan, J.C., Herron, K., Rehman, U. & Stuart, G.L. (2000). Testing the Holtzworth-Munroe and Stuart (1994) Batterer Typology, *Journal of Consulting and Clinical Psychology* **68**, 1000–1019.

[25] Hu, L. & Bentler, P.M. (1998). Fit indices in covariance structure modeling: sensitivity to underparameterized model misspecification, *Psychological Methods* **3**, 424–452.

[26] Kazdin, A.E. (2003). *Research Design in Clinical Psychology*, Allyn and Bacon, Boston.

[27] Kim, Y., Pilkonis, P.A., Frank, E., Thase, M.E. & Reynolds, C.F. (2002). Differential functioning of the beck depression inventory in late-life patients: use of item response theory, *Psychology & Aging* **17**, 379–391.

[28] Kraemer, H.C., Stice, E., Kazdin, A., Offord, D. & Kupfer, D. (2001). How do risk factors work together? Mediators, moderators, and independent, overlapping, and proxy risk factors, *American Journal of Psychiatry* **158**, 848–856.

[29] Kraemer, H.C., Wilson, T., Fairburn, C.G. & Agras, W.S. (2002). Mediators and moderators of treatment effects in randomized clinical trials, *Archives of General Psychiatry* **59**, 877–883.

[30] Lambert, M.C., Schmitt, N., Samms-Vaughan, M.E., An, J.S., Fairclough, M. & Nutter, C.A. (2003). Is it prudent to administer all items for each child behavior checklist cross-informant syndrome? Evaluating the psychometric properties of the youth self-report dimensions with confirmatory factor analysis and item response theory, *Psychological Assessment* **15**, 550–568.

[31] Loftus, G.R. (1996). Psychology will be a much better science when we change the way we analyze data, *Current Directions in Psychological Science* **5**, 161–171.

[32] MacCallum, R.C. & Austin, J.T. (2000). Applications of structural equation modeling in psychological research, *Annual Review of Psychology* **51**, 201–226.

[33] MacCallum, R.C., Roznowski, M. & Necowitz, L.B. (1992). Model modifications in covariance structure analysis: the problem of capitalization on chance, *Psychological Bulletin* **111**, 490–504.

[34] MacCallum, R.C., Wegener, D.T., Uchino, B.N. & Fabrigar, L.R. (1993). The problem of equivalent models in applications of covariance structure analysis, *Psychological Bulletin* **114**, 185–199.

[35] Masson, M.E.J. & Loftus, G.R. (2003). Using confidence intervals for graphically based data interpretation, *Canadian Journal of Experimental Psychology* **57**, 203–220.

[36] McFall, R.M. & Townsend, J.T. (1998). Foundations of psychological assessment: Implications for cognitive assessment in clinical science, *Psychological Assessment* **10**, 316–330.

[37] McFall, R.M. & Treat, T.A. (1999). Quantifying the information value of clinical assessments with signal detection theory, *Annual Review of Psychology* **50**, 215–241.

[38] Meehl, P.E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology, *Journal of Consulting and Clinical Psychology* **46**, 806–834.

[39] Miller, G.A. & Chapman, J.P. (2001). Misunderstanding analysis of covariance, *Journal of Abnormal Psychology* **110**, 40–48.

[40] Nelson, C.B., Heath, A.C. & Kessler, R.C. (1998). Temporal progression of alcohol dependence symptoms in the U.S. household population: results from the national comorbidity survey, *Journal of Consulting and Clinical Psychology* **66**, 474–483.

[41] Peeters, F., Nicolson, N.A., Berkhof, J., Delespaul, P. & deVries, M. (2003). Effects of daily events on mood states in major depressive disorder, *Journal of Abnormal Psychology* **112**, 203–211.

[42] Quesnel, C., Savard, J., Simard, S., Ivers, H. & Morin, C.M. (2003). Efficacy of cognitive-behavioral therapy for insomnia in women treated for nonmetastatic breast cancer, *Journal of Consulting and Clinical Psychology* **71**, 189–200.

[43] Raudenbush, S.W. & Bryk, A.S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods*, 2nd Edition, Sage, Thousand Oaks.

[44] Reise, S.P., Waller, N.G. & Comrey, A.L. (2000). Factor analysis and scale revision, *Psychological Assessment* **12**, 287–297.

[45] Rosenthal, R., Rosnow, R.L. & Rubin, D.B. (2000). *Contrasts and Effect Sizes in Behavioral Research*, Cambridge University Press, Cambridge.

[46] Rosnow, R.L. & Rosenthal, R. (2003). Effect sizes for experimenting psychologists, *Canadian Journal of Experimental Psychology* **57**, 221–237.

[47] Rossi, J.S. (1990). Statistical power of psychological research: what have we gained in 20 years? *Journal of Consulting and Clinical Psychology* **58**, 646–656.

[48] Scott, K.L. & Wolfe, D.A. (2003). Readiness to change as a predictor of outcome in batterer treatment, *Journal of Consulting & Clinical Psychology* **71**, 879–889.

[49] Sedlmeier, P. & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin* **105**, 309–316.

[50] Tabachnick, B.G. & Fidell, L.S. (2001). *Using Multivariate Statistics*, Allyn and Bacon, Boston.

[51] Tomarken, A.J. & Waller, N.G. (2003). Potential problems with "well fitting" models, *Journal of Abnormal Psychology* **112**, 578–598.

[52] von Eye, A. & Schuster, C. (2002). Log-linear models for change in manifest categorical variables, *Applied Developmental Science* **6**, 12–23.

[53] Walden, T.A., Harris, V.S. & Catron, T.F. (2003). How I feel: a self-report measure of emotional arousal and regulation for children, *Psychological Assessment* **15**, 399–412.

[54] Weersing, V. & Weisz, J.R. (2002). Mechanisms of action in youth psychotherapy, *Journal of Child Psychology & Psychiatry & Allied Disciplines* **43**, 3–29.

[55] Weisz, J.R., Donenberg, G.R., Han, S.S. & Weiss, B. (1995). Bridging the gap between laboratory and clinic in child and adolescent psychotherapy, *Journal of Consulting and Clinical Psychology* **63**, 688–701.

[56] Weisz, J.R., Weiss, B., Han, S.S., Granger, D.A. & Morton, T. (1995). Effects of psychotherapy with children and adolescents revisited: a meta-analysis of treatment outcome studies, *Psychological Bulletin* **117**, 450–468.

TERESA A. TREAT and V. ROBIN WEERSING