American Psychological Association

American Psychologist

© 2019 American Psychological Association
ISSN: 0003-066X

2019, Vol. 74, No. 9, 1018–1028
http://dx.doi.org/10.1037/amp0000531

Enhancing Validity in Psychological Research

David A. Kenny
University of Connecticut

Methods to increase Campbell's (1957) internal and external validity as well as Cook and Campbell's (1979) construct and conclusion validity are reviewed. For internal validity or valid causal inference, designs and methods to draw causal conclusions from nonrandomized studies are considered. Greater collaboration between the causal inference and structural equation modeling traditions would benefit both. For external validity, generalizing results, treating partners and studies as well as participants as random is strongly encouraged. For construct validity, particularly the psychological meaning of measures, multivariate models that treat measures from both overtime and dyadic data as being a combination of multiple constructs are discussed. For conclusion validity or valid statistical inference, the problem of low power when generalizability is high and the assumption of independence are discussed. Finding the truth in psychological research is a challenge, and seemingly insurmountable difficulties are often encountered. Nonetheless, persistent and diligent efforts using strategies developed by generations of methodologists should lead to scientific advancement.

Public Significance Statement
For psychological research to be applied, its conclusions must be true. Donald T. Campbell and colleagues developed four different ways to assess the validity of scientific research. In this article, I discuss work done by myself and my colleagues on enhancing these four different types of validity.

Keywords: validity, quasi-experiment, causation, power, meta-analysis

In a 1957 *Psychological Bulletin* article (Campbell, 1957), Donald T. Campbell urged psychologists to consider the different ways that the conclusions from psychological research might be true, and he coined the terms *internal* and *external validity*. These ideas were later elaborated in the better-known little book that Campbell wrote with Julian Stanley (Campbell & Stanley, 1963). Later, in their book *Quasi-Experimentation: Design & Analysis Issues for Field Settings* (Cook & Campbell, 1979), Thomas D. Cook and Campbell introduced construct and conclusion validities, and in 2002, with Will Shadish, extended that discussion in *Experimental and Quasi-Experimental Designs for Generalized Causal Inference* (Shadish, Cook, & Campbell, 2002). In this article, I discuss how my collaborators and I have strived to find ways to enhance each of these validities.

For internal validity or valid causal inference, I discuss the use of structural equation modeling (SEM) to test causal theories when one is unable to randomly assign participants to conditions. There is a discussion of recent advances in causal inference and mediation. For external validity or generalizing results, I discuss the importance of generalizing not only across participants but also across targets or stimuli and studies themselves. For construct validity or the meaning of measures (as well as causes, effects, and settings), I discuss how psychological measures typically refer to not only the intended construct but also to additional constructs. For instance, a dyadic measurement refers to the respondent, the other member of the dyad, and their relationship; in longitudinal studies, a measurement refers to a stable trait, a slow-changing trait, and a state. Finally, conclusion validity refers to the correctness of the statistical conclusions derived from research. I examine the assumption of independence in statistical models as well as dealing with low power when trying to achieve greater generality.

Editor's note. David A. Kenny received a 2019 APA Award for Distinguished Scientific Contributions. This article is based on an invited presentation at the 127th Annual Convention of the American Psychological Association, held August 8–11, 2019, Chicago, Illinois.

Author's note. David A. Kenny, Department of Psychological Sciences, University of Connecticut.

I thank Betsy McCoach, Emil Coman, Stephen West, Charles Judd, and Shelley Riggs for comments on an earlier version of this article.

Correspondence concerning this article should be addressed to David A. Kenny, Department of Psychological Sciences, University of Connecticut, 406 Babbidge Road, Unit 1020 Storrs, CT 06269-1020. E-mail: david.kenny@uconn.edu

Although the article discusses some very technical issues, I have striven to do so in a nontechnical fashion with references to more technical sources.

Internal Validity

Causation: Going Beyond the Randomized Experiment

The backbone of psychology, rightly so, rests on randomized studies, analyzed by analysis of variance (ANOVA). As a first-year graduate student in 1969, I labored in the calculation of sums of squares and mean squares using a Monroe calculator. A large repeated-measures ANOVA took a day to calculate. I was also fortunate to teach ANOVA at Harvard University for 5 years in the 1970s, 1 year with the statistician Paul Holland. Despite the beauty and elegance of randomized experiments being analyzed by ANOVA, that strategy for estimating causal effects is limiting. The major limitation, to my mind, is the requirement to manipulate the causal variable.

Certainly, being able to randomly assign participants to conditions is an excellent way to establish internal validity, but it is not the only way. I shudder when I hear or read the following oft-repeated statement, “I am unable to draw a causal conclusion because I did not randomize participants to conditions.” If causal conclusions were limited to only variables that can be manipulated, then the scientific study of causal effects would be precluded for many key psychological variables, for example, self-esteem, childhood abuse and malnutrition, birth order, and ethnicity. A major reason for writing my 1979 book, *Correlation and Causality*, was to illustrate how it was possible to measure causal effects without randomization. By not randomizing, there are what Campbell called *threats to validity*. The key threat in nonrandomized studies is that of selection: Persons who received the intervention differed on some variable that also causes the outcome variable. Charles Judd and I (Judd & Kenny, 1981a) referred to this type of variable as the *assignment variable*, and the more contemporary term is *confounder*. One of Campbell’s goals in developing quasi-experimental designs was to find a way around the problem of confounding. Different assumptions are made about the confounder for each of the quasi-experimental designs. For the regression discontinuity design, the assumption is linearity, whereas for the nonequivalent control group design and interrupted time series designs, more elaborate assumptions about the confounder are made (Judd & Kenny, 1981a; Kim & Steiner, 2019). Despite not randomizing, these quasi-experimental designs yield estimates that are typically very close to the estimates that would have been obtained in a randomized experiment (Wong, Steiner, & Anglin, 2018).

For observational studies in which all observations are all measured at the same time, commonly called a *cross-sectional study*, it becomes a more difficult task to remove the effects of confounding variables. The standard advice is to ascertain the variables that might be confounders, to measure them, and control for them in the statistical analysis or in the design of the research. However, in practice, this is not so easy. First, one needs to be careful about what variables are controlled, as sometimes controlling for a supposed confounder only makes things worse.[1] Second, some confounders can be difficult to measure. Consider the classic example of smoking causing lung cancer. Perhaps there is a gene that causes people to smoke and to contract lung cancer. How can we measure this gene if we do not even know what it is? The confounder is an unmeasured variable. There are some sophisticated strategies, but not widely known in psychology, to control for unmeasured confounders, for example, instrumental variables and the front-door method (see Pearl & MacKenzie, 2018). Moreover, several different sources (that is, Abelson’s, 1995 MAGIC criteria; Shadish et al., 2002) discuss more conventional strategies to enhance internal validity in observational studies.

To draw causal conclusions from any study requires a detailed inquiry. Even randomized experiments may have issues of missing data and noncompliance (i.e., nonadherence), which compromise causal conclusions. In nonexperimental research, many issues—for example, confounding, measurement error, and criteria for inclusion—require careful consideration and analysis. As Campbell and I discussed in our 1999 book (Campbell & Kenny, 1999), one needs to be one’s own worst critic. Elaborate and rule out plausible rival hypotheses by logic, further analyses, or new data. Another way to rule out rival hypotheses is to conduct sensitivity analyses or “what if” analyses. One example of a sensitivity analysis is to determine how strong the effect of the confounder would have to be to render an estimated causal effect to become zero.

Diagrams and Structural Equation Modeling

A tool that psychologists regularly employ to measure causal effects in nonrandomized research is a path diagram. As shown in Figure 1, a path diagram contains representation of causal effects as arrows from a cause to an effect, for example, from Z to R. A path diagram normally includes several different outcome variables (Z, R, and T) as well as confounding variables (Z being a confounder of the R and T relationship) and mediational chains (Z being a mediator of the relationships between X and Y with R and T). The standard convention in path diagrams is that unmeasured variables are denoted by circles and measured variables by boxes. Correlations between variables that are not caused by

[1] There are what are called *colliders* (Pearl & MacKenzie, 2018), which, when controlled, increase and not decrease the bias in the estimated causal effect.

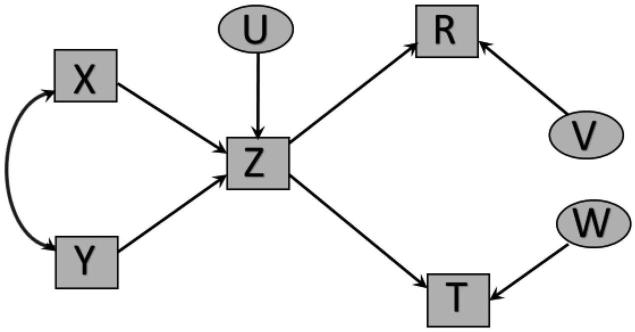

Figure 1. Illustration of a path diagram.

any other variable in the model (i.e., exogenous variables) are represented by a curved line (e.g., between *X* and *Y*). The path diagram summarizes the formal set of causal equations and, in fact, does so better than the equations.

Once the paths are known, the researcher may[2] be able to estimate the paths. With those paths, there are two key ways to falsify the model. First, it may be the case that when the model is estimated, the key path of interest is estimated (e.g., the path from *Z* to *R*) to be very small and hardly different from zero. Second, the model may imply that certain paths are zero. In Figure 1, there are no direct causal paths from *X* and *Y* to *R* and *T*. Verification that these specified zero paths are in fact zero is sometimes a much stronger confirmation of model than finding that a specified path is in fact nonzero.

Psychologists interested in causal models have been greatly aided by the development of SEM programs. Some SEM programs (i.e., AMOS [Arbuckle, 2015] and EQS [Bentler, 1992]) accept as input a path diagram. These models can be used to estimate very complicated models with several different outcome models for different groups of participants (e.g., treated and controls), the presence of measurement error in variables, and even latent variables that interact (Kenny & Judd, 1984). Initially, these SEM models were explicitly causal models. Unfortunately, somewhere along the line, many (but not all, myself included) psychologists abandoned or avoided the concept of causality. The term *structural* started to replace *causal*: Models were “structural” models and paths were “structural” paths. This dropping of the term *causal* from SEM models has led to the unfortunate practice of researchers saying, “I am unable to draw a causal conclusion because I did not randomize participants to conditions.”

There has been a revival of causal thinking, largely outside of psychology, centered primarily in epidemiology, and that movement is referred to as *causal inference*. A good introduction to this approach is the *Book of Why* by Judea Pearl and Dana MacKenzie (Pearl & MacKenzie, 2018). These researchers used directed acyclic graphs (DAGs), not path diagrams. Path diagrams are for linear models, whereas DAGs permit nonlinearities (e.g., models in which all variables are dichotomies or when variables interact). Figure 2 presents the corresponding DAG for the path model drawn in Figure 1. Besides DAGs allowing for nonlinearities and path models not, there are three other differences between the two.

The first difference is that using boxes for measured variables and circles for latent variables is not usually done for DAGs. This is just a stylistic difference, and I think it is helpful to make that designation, which is why I have done so in the Figure 2 DAG.

The second difference is that a path diagram allows for causal variables to be correlated, and they denote such by a curved lines. Note the curved line between variables *X* and *Y* in the path diagram in Figure 1. DAGs traditionally do not allow any curved lines. In most cases, the source of the correlation is a potential unmeasured confounder. For the DAG in Figure 2, the curved line between *X* and *Y* in Figure 1 is replaced with a confounder *C*, which has paths to both *X* and *Y*. I think this is a useful practice that should be adopted in path diagrams. Note that the DAG in Figure 2 makes it clear that *X* and *Y* mediate the causal effect of the confounder, *C*, on *Z*, *R*, and *T*. Almost always, SEM researchers fail to realize this hidden assumption in their causal models, but adding *C* to the model makes that assumption more obvious.[3]

The third difference is that DAGs do not ordinarily have disturbances for effects. Path diagrams have disturbances, which represent all of the other causes of a variable besides those specified in the model. A disturbance is similar, though not exactly the same, as an error term in a conventional linear model. Very often in causal modeling, a key assumption is that the disturbance is uncorrelated with all of the causes of an effect (Kenny, 1979). This is explicitly shown in the path diagram in Figure 1: The disturbance for *Z*, denoted as *U*, is uncorrelated with *X* and *Y*, the two causes of *Z*. The assumption is implicit in the DAG in Figure 1, but nonetheless there are advantages to making it explicit. For instance, in Figure 1, one might want to correlate *V* and *W*. It might be the case that the two variables are measured by a common method, and it would seem to be advisable that there is correlated disturbances between the two measures.

Although, historically, most psychologists have backtracked on the goal of drawing causal inferences from

[2] The determination of whether the paths in a given model can all be uniquely estimated is called *identification* in structural equation modeling. It is commonly, but mistakenly, assumed that all paths in a model need to be identified to be able to estimate a model (Kenny & Milan, 2012). For instance, in the model in Figure 2, the paths from *C* to *X* and *Y* are not identified, but conceptually this is unimportant.

[3] Colin Leach and I realized these points in our discussions about multicollinearity. Together, we realized that sometimes it might be more plausible to assume that *C*, not *X* or *Y*, causes *Z* (see Figure 2).

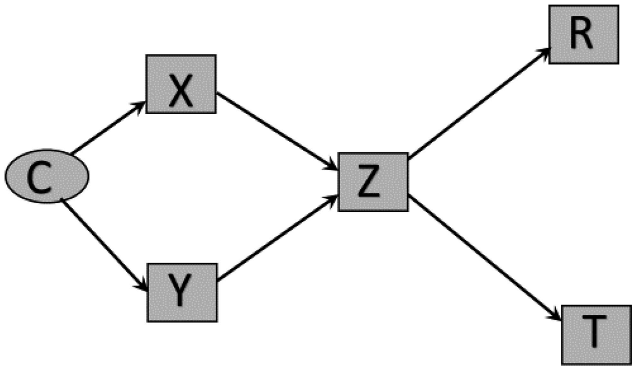

Figure 2. The path diagram in Figure 1 redrawn as a directed acyclic graphs (DAG).

nonexperimental studies, they can, and should, become collaborators with those in the causal inference camp, especially if they seek to draw causal conclusions. Psychologists can offer sophisticated tools for handling measurement error and other complications through the application of SEM programs. Granted, these programs are mainly, but not exclusively, for linear models. The focus on dichotomous measures such as life and death is understandable in epidemiology, but many psychologists often use scales for measurement. I would hope that that there would be increasing collaboration between those who use SEM and the causal inference camp, something that has already begun (e.g., Bollen & Pearl, 2013; Muthén & Asparouhov, 2015).

Mediation

For this section, a simple mediational model is assumed. One starts with a variable *X*, which may or may not be manipulated, and two outcome variables. One outcome is the mediator, *M*, which is assumed to be caused by *X*, and the other is the outcome, or *Y*, which is assumed be caused by both *X* and *M*. Figure 3 illustrates the mediational model as a DAG.

A key part of most causal models is the mediational part, which leads to interest in the topic of mediation. I coauthored two articles, both of which I was deservedly the second author, with Charles Judd (Judd & Kenny, 1981b) and Reuben Baron (Baron & Kenny, 1986), which attempted to outline how to conduct a mediational analysis. Mediation analysis is causal analysis: A researcher assumes that part of the causal effect of one variable, *X*, on another, *Y*, is due to the causal effect of a third variable, *M*. Unfortunately, the published version of the oft-cited 1986 article does not emphasize enough the causal assumptions of mediation. The original article that we submitted did, but in the editorial process, those explicit statements of the causal assumptions were cut, and we were told to cite the earlier 1981 article. Unfortunately, not many of the readers of 1986 article also read the 1981 article, and all too often, mediational articles make no attempt to justify the causal assumptions underlying their model. Those causal assumptions are as follows. When *X* is randomized, the causal assumptions in a mediational analysis are as follows:

1. There exist no confounding variable that causes both *M* and *Y*.
2. There is no measurement error in *M*.
3. The variable *Y* does not cause *M*.

If *X* is not randomized, then the following must also be assumed:

1. There exist no confounding variable that causes both *X* and *M* or both *X* and *Y*.
2. There is no measurement error in *X*.
3. The variable *X* is not caused by *M* or *Y*.

Researchers need to do the hard work to justify these assumptions and probe the consequences if the assumptions were violated. An excellent introduction to causal inference with observational data is given in Rohrer (2018).

As discussed earlier for quasi-experimental designs, design features assist in meeting causal assumptions, and a key design feature is the timing of measurement. For instance, in Morse, Calsyn, Allen, and Kenny (1994), we examined how assignment to an intensive case management (*X*) increases the number of contacts with housing agencies (*M*), which, in turn, increases the number of days stably housed per month for homeless individuals (*Y*). Persons were randomly assigned to intensive case management, and then for the next 9 months, the number of housing contacts was measured, and then in the next 7 months, stable housing was measured.

More commonly, and more problematically, a cross-sectional design is used to measure mediation. A key article by Maxwell and Cole (2007) showed how such a practice

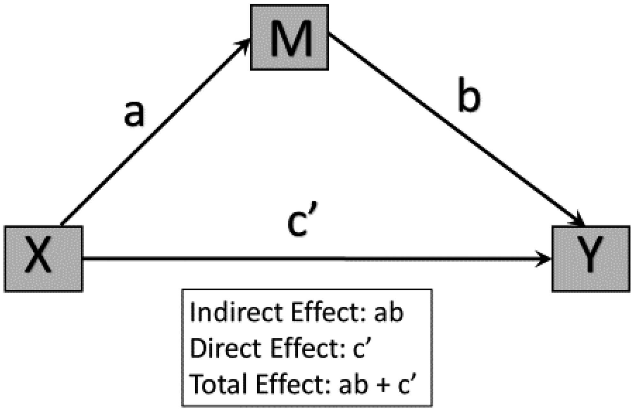

Figure 3. Three-variable mediation path diagram.

can be problematic. Certainly, an uncritical use of a cross-sectional design is problematic, but not always fatal, a point alluded to by Maxwell and Cole (see p. 40). Consider the study by Riggs, Cusimano, and Benson (2011). They were interested in the mediating effect of attachment styles of both oneself and one's relationship partner, which mediated the causal effect of childhood abuse on relationship satisfaction. They measured all of these variables from 155 heterosexual couples at one time. The authors presented evidence of the validity of retrospective measures of childhood abuse (see pp. 134–135). There is the worry that relationship satisfaction might affect attachment style, but likely, the preponderance of the causation goes from attachment to satisfaction. Thus, in this case, a cross-sectional mediation analysis is not as problematic as it might seem.

When *X* is experimentally manipulated, the key worry is bias in estimating the effect from *M* to *Y*, Path *b* in Figure 3. As discussed in Fritz, Kenny, and MacKinnon (2016), when *X* is a manipulated variable, there are two likely sources of bias. One is an unmeasured confounder that causes both *M* and *Y*. The other is measurement error in *M*. As Fritz et al. showed, very often, but not always, these two biases work in the opposite direction. Confounding tends to artificially inflate the *b* path, whereas measurement error deflates it. These two biases likely do not exactly cancel each other out, but things are not as bad as they seem. Also ironically, just controlling for one of the biases might create more bias than not doing anything.

My last point about mediation concerns an anomaly that has been pointed out by numerous authors, perhaps first by the eminent statistician David Cox (1960) and reviewed in a 2014 article that I wrote with Charles Judd (Kenny & Judd, 2014). Imagine a mediational model in which there is complete mediation, that is, $c' = 0$ (see Figure 3). In such a case, the total effect, the effect of *X* on *Y* ignoring *M*, is equal to the indirect effect, the effect of *X* on *Y* due to *M*, both equaling *ab*. However, typically substantially fewer participants are needed to have sufficient power (the chance of finding a significant effect), sometimes 20 times less, in the test of the indirect effect than in the test of the total effect. This result could be used to measure a causal effect using the indirect effect with more precision than using the effect itself. If it is known that *M* completely mediates the *X* to *Y* relationship, and the effects from *X* to *M* and from *M* to *Y* can be estimated without bias (these being very strong assumptions), it would be possible to obtain much more precise measures of causal effects using the indirect effect than the total effect. As an example, in a reanalysis of a large HIV prevention study, Eaton, Kalichman, Kenny, and Harel (2013) showed that an intervention to reduce risky sexual practices (the mediator) did lead to a beneficial effect of the intervention on HIV rates when measured by the indirect effect that was not readily apparent when only the total effect of the intervention was estimated.

External Validity

It is not generally understood that the key purpose of significance testing to determine whether a result (e.g., a mean difference between conditions) would replicate if the study was repeated using a different set of participants. Certainly, this is an important goal, but psychologists often seek to generalize over more than just participants. In particular, it is important to generalize across targets (to be explained later) and across studies themselves.

Generalization Across Targets

Very often in research, there is a second factor, besides participants, across which psychologists seek to generalize. Consider the following examples:

> In a study of word recognition as function of the number of syllables, a cognitive psychologist wants to generalize across words as well as participants.
>
> A personnel psychologist, who is studying the salaries offered to female and male job applicants, wants the result to generalize beyond the particular male and female applicants used in the study.
>
> A clinical psychologist, who is studying the efficacy of two different types of psychotherapy, seeks to generalize not only across clients but also therapists.
>
> A social psychologist, who is looking at implicit biases of White versus Black faces, wants to generalize beyond the particular faces used.

All of these examples show that in many areas of psychology, there is a second factor that needs to be generalized over besides participants. I refer to the second factor as *targets*, which makes syllables, applicants, therapists, and faces the *targets* in the above examples.

In an *Annual Review of Psychology* chapter that I wrote with Charles Judd and Jake Westfall (Judd, Westfall, & Kenny, 2017), we explored the implications of generalizing across a second factor. To be able to generalize across targets, as well as participants, targets must be treated a random variable in the analysis. In the case of equal cell sizes and no missing data, these data could be analyzed by a repeated measures ANOVA. However, to allow for likely complications in the analysis, almost always required are mixed model analyses that are extensively reviewed in Judd et al. chapter.

Let me call the experimental variable of interest *treatment* and presume that it has two conditions. Judd et al. (2017) reviewed a number of different design options that are based on three different choices:

Is each participant in both treatment conditions or just one (between vs. within design)?

Is each target in both treatment conditions or just one?

Does each participant respond to all possible targets or to just a subset?

The full set of possible designs is detailed in Judd et al.[4]

When there is a second random factor, an often-ignored, yet crucial, consideration is the number of targets. All too often, an insufficient number of targets are used in studies with two random factors. As Judd et al. (2017) showed, if there are too few targets, even if there is a very large number of participants, the study might have still a low chance of finding a meaningful result. Only by having a sufficient number of targets, as well as participants, can meaningful results be obtained. One way to improve power is to internally replicate the design. Consider a design in which the targets are male and female faces, in which gender is the experimental variable. One design would be to have each participant judge all the faces, say eight male and eight female. An alternative, and far better, design would be to have replicates of this design. In each replicate, the participants judge 16 different faces. Jake Westfall has a website to aid in making that choice: http://jakewestfall.org/two_factor_power/.

Generalization Across Studies

Repeated investigations of the same phenomenon typically yield effect sizes that vary more than one would expect from sampling error alone, something called *heterogeneity*. As Judd and I (Kenny & Judd, 2019) and others (e.g., McShane, Tackett, Böckenholt, & Gelman, 2019) have documented, such variation is even found in close replication studies. The usual explanation for heterogeneity is that there are what are called *hidden moderators*: It should be possible to specify what these moderators are and use them to explain the heterogeneity. However, attempts to find these moderators are typically unsuccessful. For instance, Klein et al. (2018) concluded the following: “We observed some heterogeneity between samples, but a priori predictions . . . and prior findings . . . were minimally successful in accounting for it” (p. 77). Similar attempts to discover moderators of heterogeneity (e.g., Linden & Hönekopp, 2019) have been repeatedly unsuccessful. It appears that heterogeneity cannot be eliminated by specifying the anticipated moderators. Judd and I speculate that at the heart of heterogeneity is either a fundamental or practical randomness of effects.

The presence of heterogeneity undermines the informativeness of a single large-*N* study. Certainly, a single small-*N* study is problematic for several reasons. First, a small-*N* study has low power. Second, a published small-*N* study likely is an overestimate of effect size because of publication bias. However, a single large-*N* study is not the solution. Given heterogeneity and the conventional assumption of a normal distribution of effects, a large-*N* study can lead to a nontrivial chance of finding a results in the wrong direction, that is, one opposite in sign to the average effect size, than a small-*N* study. This surprising conclusion rests on the traditional assumption that the distribution of random effect sizes is normal. Judd and I (Kenny & Judd, 2019) argue that a better strategy are multiple moderate-*N* studies, a practice commonly practiced in multiple replication projects (e.g., Klein et al., 2014, 2018).

There is strong circumstantial evidence that the distribution of effects might not be normal but may have a lower limit of zero (assuming the average effect is positive). In several investigations (e.g., Klein et al., 2014, 2018), studies whose average effect size is near zero have little or no heterogeneity. When there is heterogeneity, the average effect size is nonzero. As Judd and I discuss (Kenny & Judd, 2019), such a result is consistent with view that distribution of true effect sizes is positively skewed with a lower bound of zero.

Although non-normal effect sizes are statistically messy, they have the benefit of removing a major ethical dilemma. Imagine a beneficial intervention to reduce posttraumatic stress disorder (PTSD), but there is some heterogeneity: The beneficial effect is not always the same, sometimes being large and sometimes not so large. If the average effect is not too large and there is a fair bit of heterogeneity, assuming a normal distribution, then sometimes the “benefit” could be negative, meaning that the program could be harmful and increase PTSD. Would it then be ethically acceptable to recommend a program that is helpful on average but sometimes harmful? Assuming the distribution of effects has a lower bound of zero removes this ethical dilemma.

Construct Validity

A key idea in Campbell’s work is that underlying a given measurement are other constructs besides the particular construct that was intended to be measured. For instance, if a researcher wants to measure how depressed a child is, a parent’s report of the child’s depression would measure something else besides the child’s depression. To determine how much variance is due to these different constructs, there needs to be multiple measurements. For instance, to separate trait and methods requires measuring each trait by each method to produce what Donald Campbell and Donald Fiske called the *multitrait-multimethod matrix* (Campbell & Fiske, 1959). In two disparate areas, I have proposed that it is beneficial to separate a measure into its different compo-

[4] There is one more design, the counterbalanced design, that does not neatly fit into the typology.

nents, and to accomplish this, there needs to be multiple measurements.

Longitudinal Measurement

A classic question in psychology is whether something, for example, anxiety, is a trait (i.e., essentially unchanging) or a state (totally changing). The position that I took with Alex Zautra in two articles (Kenny & Zautra, 1995, 2001) was that most psychological constructs are simultaneously both traits and states. Just because a measure is called *trait anxiety* does not necessarily mean that all its variance is trait variance. In our STARTS model, Alex and I partitioned the total variance of a measure into three sources. The first is a component that does not change over time, which we called a *stable trait*; the second is a slowly changing component, called the *autoregressive trait*; and the final is a component that totally changes at each measurement and is called a *state*. To accomplish such a partitioning, at least four waves of measurement are desirable, but more is preferable, and ideally a large sample size.

Brent Donnellan, I, and several other colleagues performed a STARTS analysis of self-esteem (Donnellan, Kenny, Trzesniewski, Lucas, & Conger, 2012). We had data from the Iowa Youth and Families Project (Elder & Conger, 2000), in which 451 people were measured on the Rosenberg Self-Esteem Scale (Rosenberg, 1965). Participants were first measured at Age 13 and then measured nine more times, ending when they were 32 years old. We found that the stable trait accounted for 35% of the total variance, the autoregressive trait accounted for 49%, and the state, 16%. The rate of change in the autoregressive trait (i.e., its correlation) was .89 in a year (from Ages 21 to 22), and that rate was deaccelerating from year to year, with there being less change as people aged. Using STARTS, we forecasted the age at which self-esteem first emerged and estimated it to be at about 11.7 years of age, and before that, it therefore makes no sense to refer to the "child's self-esteem."

In a subsequent replication study by Wagner, Lüdtke, and Trautwein (2016), a total of 4,532 Germans, Ages 20 to 30, were measured at six waves, each 2 years apart. The self-esteem measure was the Self-Descriptive Questionnaire (Marsh, 1992). The investigators conducted separate analyses by gender and level of depression. They found that the stable trait accounted for 44% of the total variance, the autoregressive trait accounted for 42%, and the state, 14%. The rate of change in the autoregressive trait was .85 a year (from Age 25 to 26), and that rate was deaccelerating from year to year. Using STARTS, they forecasted the age at which self-esteem first emerged and estimated it to be at about 12.4 years of age. Considering that are the differences between the studies in nationalities, measures, and age range, there is a remarkable consistency of these two STARTS studies.

So is self-esteem a trait? This is the wrong question to ask. It is not one thing but, rather, a mix of three things. Moreover, the smallest part in terms of variance, state, might be the most important component. Note that Wagner et al. (2016) concluded that "females showed more state variability than males. Individuals with higher levels of depressive symptoms showed more state and less autoregressive trait variance in self-esteem" (p. 523). They went on to suggest that perhaps those who are more reactive to environmental factors may be particularly vulnerable to depressive episodes.

Dyadic Measurement

My substantive research focuses on person perception. Consider two coworkers, Sally and Sue, and how much Sally likes Sue is measured. Different types of psychologists focus on different aspects of this measure of interpersonal attraction: A clinical psychologist might want to know whether Sally is the sort of person who is trusting and likes everyone that she meets. An evolutionary psychologist might be interested in Sue's physical appearance and determine whether she is the type of person everyone likes. A relationship researcher might want to know if Sally particularly likes Sue. Underlying a "simple" dyadic measurement lurks many different components. It would be a mistake to treat the measurement as if only just one of these components mattered. A complete analysis of interpersonal perception requires consideration of multiple components.

Building on earlier work I did in 1979 with Rebecca Warner and Michael Stoto (Warner, Kenny, & Stoto, 1979), Lawrence La Voie and I (Kenny & La Voie, 1984) proposed the social relations model as a way to conceptualize dyadic measurements. In this model, a dyadic perception is assumed to be a function of the perceiver, target, and relationship components. Sally's liking of Sue is due to

> a perceiver effect: the tendency for Sally to like or dislike everyone; referred to as "how choosy" the person is (the larger the perceiver effect, the *less* choosy is the person);
>
> a target effect: the extent to which Sue is generally liked or disliked by others; referred sometimes as how *popular* the target is; and
>
> a relationship effect: the extent to which the Sally particularly likes or dislikes the Sue controlling for how much Sally likes others and how much Sue is liked by others.

As with the STARTS model, to be able to separate these different components, there needs to be multiple measurements. One way of accomplishing this is to have Sally, Sue, and their coworkers state how much they all liked one another. Alternatively, a family psychologist might study liking in families and ask all family members how much they like one another. The perceiver effect would be reflected by Sally's tendency to like all of her coworkers; the target effect would be reflected by the tendency for Sue to be liked by all of the other coworkers; and the relationship effect by Sally especially liking Sue, technically the interaction of Sally with Sue.

In my 2020 book *Interpersonal Perception: The Foundation of Social Relationships*, I explored these different components. Here, I report variance partitioning results from nine different studies of interpersonal attraction of people who know each other well (Kenny, 2020, Table 5.2). These nine round-robin studies are highly diverse, ranging from attraction ratings of the University of Texas marching band to fifth and sixth grade children in Nijmegen, the Netherlands. The largest component is the relationship effect, which explains 64% of the total variance. There is, indeed, no accounting for taste. These relationship effects also tend to be reciprocal, with that correlation being .58 (Kenny, 2020, Table 5.5)—that is, Sally's unique liking of Sue is highly correlated with Sue's unique liking of Sally.

Target variance is the tendency for some people to be generally liked and other people to be not liked so much. This component explains 20% of the total variance. Note that for romantic attraction in speed dating, there is a dramatic increase in target variance, about doubling. The source of the jump is that physical attractiveness matters much more in romantic than in friendship relationships.

The smallest source of variance is perceiver variance, which explains 16% of the total variance. Some of this variance is due to individual differences, in that some people are choosy about whom they like and others are not. Another explanation of this source of variance is that it reflects the perceiver's feeling about the group in general. If the perceiver likes being the group, he or she would tend to like the members of that group.

Conclusion Validity

One conducts a statistical analysis, and from that analysis, one computes a *p* value, a confidence interval, or an equivalence test (Schuirmann, 1987). I focus on significance testing, but the statements in the section apply to other forms of model testing, including Bayesian methods. Conclusion validity focuses on two parts of significance testing. First, in planning the study, we want to know whether the study is large enough to yield interpretable results. Conventionally, the question is power or the probability of rejecting the null hypothesis. Another part of conclusion validity concerns the truth of assumptions that are made in the statistical model. Here, the focus in is on the assumption of independence.

Power

Earlier, I discussed two different ways to increase external validity: generalizing across targets and studies. Doing so often results in a low-power study, a study with little chance of finding a statistically significant result. Consider a totally hypothetical study to show that female faces are generally seen as being more likable than male faces, the effect size of *d* being 0.5. Consider a study to investigate that hypothesis, and 128 participants judge the same eight male and eight female faces. Using the standard assumptions about the variance of effects made by Judd et al. (2017; see Figure 2) and treating only rater (not face) as a random variable, there is a 92% chance of finding a positive significant result confirming the hypothesis that female faces are more likable, but there is about a 2% chance of finding a significant result that men are more likable. However, correctly treating face as a random variable results in a power of just .38. Increasing generality across both faces and participants results in a loss of power. Some of the power can be recouped by using replicates. That is, if there are four sets of 16 faces, each judged by 32 participants, the resulting power would be an impressive .90, with almost no chance of finding that male faces are more likable. The bottom-line result is that power is necessarily less if targets are treated as random. Additionally, allowing for heterogeneity of effects across studies also results in a loss of power and precision (Kenny & Judd, 2019). Though discouraging, it seems all but inevitable that broader generalizations come with cost of less power and precision.

We are then faced with a dilemma: Do we want research to be more generalizable or do want to obtain statistically significant results? One way to solve the dilemma is to make psychological research collaborative. When I went to graduate school, the model of the ideal psychologist was someone who worked alone. One might have graduate students doing some of the studies, but the classic paradigm was the solitary scientist. This paradigm has totally changed, and that change needs to accelerate. Increasingly, there are signs of collaborative projects. Among them are the Many Labs project (e.g., Klein et al., 2014, 2019); Registered Replication Reports, or RRR (Simons, Holcombe, & Spellman, 2014); and the Psychological Science Accelerator (Moshontz et al., 2018). Single, relatively low-powered studies are combined to create a more precise estimate of an effect.

Another idea is to incorporate the results of past research into analysis. Before discussing how this can be done, let me discuss an error often made in data analysis. We often make strong assumptions, known to be false, that some factor has no effect, because it is unclear exactly what value the alternative takes. Let me give some examples, the first three of which were previously discussed:

1. It is assumed the effect size for a given phenomenon has zero heterogeneity (i.e., does not vary from study to study), because if it were heterogeneous, we do not know how heterogeneous it would be.

2. In estimating causal effects, we assume that there is no confounding variable, because if we allowed for a confounder, we would have to know how large the effect of the confounder would be.

3. In estimating a causal effect, we assume that the causal variable has no measurement error, because if we allowed for measurement, error it is unclear how much measurement error there would be.

4. One knows that the missing data in one's experiment are not random, but one does not know just how nonrandom they are and so one just assumes that they are random.

As is seen in each case, it is known that some parameter is nonzero, but because we do not know exactly what the parameter does equal, we quite mistakenly just assume that it is zero. This might be called the *zero fallacy*. What are we to do?

There is an approach where one can allow for a nonzero value for a parameter but allow that parameter to have a distribution of possible values. That distribution could even be updated by the data. Bayesian methods do allow for such possibilities, and these methods are likely going to become more prevalent in the future to avoid the zero fallacy.

The Independence Assumption

In 1986, Charles Judd and I discussed a key assumption in data analysis: the assumption of independence, which is that the errors in the model are uncorrelated, making replications independent (Kenny & Judd, 1986). Researchers often worry about non-normality and unequal error variance, but typically ignore the more important assumption of independence. As Judd and I showed, violating the assumption of independence can sometimes lead to too many Type I errors, but other times to too many Type II errors. That is, by considering the nonindependence in the data, a researcher might find a statistically significant effect that he or she would have missed had he or she falsely assumed independence.

Most of my work on the nonindependence problem was for determining how to study groups and dyads. When I began in the late 1970s, the focus in the literature was testing the null hypothesis that members of the same group had independent responses. If they were independent, then "person" could be the unit of analysis. If not, then "group" or "dyad" became the unit. This approach made sense then because multilevel modeling tools were not developed. However, now that we have multilevel modeling tools, the more contemporary approach to the problem of nonindependence is to allow for the simultaneous effects at multiple levels of analyses, instead of selecting person or group as the unit of analysis.

Initially, I worked on the problem of "handling" nonindependence to yield proper significance test results of effects. Eventually, I realized that the nonindependence that occurs in dyadic and group data is not a problem to be finessed or eliminated but, rather, something inherently interesting to be studied. One way to study nonindependence is the earlier-described social relations model, which is a complicated model of dyadic nonindependence. That model simultaneously measures five different sources of nonindependence. It, however, requires complicated designs; for instance, each person in the group must provide a rating of every other member in the group, that is, a round-robin design. A much more common design is one in which each member of the group or dyad provides a single set of measures. One model that Deborah Kashy and I (Kashy & Kenny, 2000) have developed for such a design is the highly popular actor–partner interdependence model. In that model, there are two variables, *X* and *Y*, for each member of the dyad or group. A person's *Y* is assumed to a function of the person's own *X*, the actor effect, and the other person's *X*, the partner effect. That model has been applied to a myriad of different dyads: therapist–client, married and dating couples, coworkers, coach–athlete, parent–child, business partners, and friends. It essentially models nonindependence by treating the two *X*s as confounders that bring about that nonindependence.

One example of an actor–partner interdependence model study is by Klumb, Hoppmann, and Staats (2006), who studied the effect of housework on stress levels measured using cortisol levels. They studied 52 German dual-earner couples with at least one child. The actor effect was found to be positive: Doing more housework yourself is associated with higher stress levels. Interestingly, the partner effect was negative and about the same absolute size as the actor effect, indicating that the more housework was done by the partner, the lower the person's stress levels. The pattern of results suggest a contrast effect (Kenny & Cook, 1999; Kenny & Ledermann, 2010): One feels less stress if one does less housework than one's partner. Very often, the relative size of actor and partner effects is the most important part of a dyadic study.

Conclusion

This article highlights my work on enhancing validity. I have been privileged to stand on the shoulders of a true giant, my graduate advisor, Donald Campbell. It should already be clear that much of that work was done in collaboration. I have been fortunate to have many excellent collaborators throughout my career, and I want to thank them all. Particular thanks go to Charles Judd, with whom I have more publications.

The search for the truth is not easy. Donald Campbell often said that the common metaphor of ignorance as darkness and knowledge as a shining light is backward. Very often, finding the truth requires a difficult journey, following false leads and spinning one's wheels and getting nowhere. Success often comes not from some brilliant insight but through repeated failed efforts. The metaphor that he was fond of using is that scientists were like rats running in a maze with many blind alleys. Eventually, despite all of the difficulties, the truth is discovered, not so much by brilliant insights but by hard work and many failures.

References

Abelson, R. P. (1995). *Statistics as principled argument*. Mahwah, NJ: Erlbaum.

Arbuckle, J. L. (2015). *IBM SPSS AMOS 24 user guide*. Retrieved from http://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/24.0/en/amos/Manuals/IBM_SPSS_Amos_User_Guide.pdf

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. http://dx.doi.org/10.1037/0022-3514.51.6.1173

Bentler, P. M. (1992). *EQS: Structural equations program manual*. Los Angeles, CA: BMDP Statistical Software.

Bollen, K. A., & Pearl, J. (2013). Eight myths about causality and structural equation models. In S. Morgan (Ed.), *Handbook of causal analysis for social research* (pp. 301–328). Dordrecht, the Netherlands: Springer. http://dx.doi.org/10.1007/978-94-007-6094-3_15

Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin, 54,* 297–312. http://dx.doi.org/10.1037/h0040950

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

Campbell, D. T., & Kenny, D. A. (1999). *A primer on regression artifacts*. New York, NY: Guilford Press.

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Chicago, IL: Rand McNally.

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

Cox, D. (1960). Regression analysis when there is prior information about supplementary variables. *Journal of the Royal Statistical Society Series B. Methodological, 22,* 172–176. http://dx.doi.org/10.1111/j.2517-6161.1960.tb00363.x

Donnellan, M. B., Kenny, D. A., Trzesniewski, K. H., Lucas, R. E., & Conger, R. D. (2012). Using trait-state models to evaluate the longitudinal consistency of global self-esteem from adolescence to adulthood. *Journal of Research in Personality, 46,* 634–645. http://dx.doi.org/10.1016/j.jrp.2012.07.005

Eaton, L. A., Kalichman, S. C., Kenny, D. A., & Harel, O. (2013). A reanalysis of a behavioral intervention to prevent incident HIV infections: Including indirect effects in modeling outcomes of Project EXPLORE. *AIDS Care, 25,* 805–811. http://dx.doi.org/10.1080/09540121.2012.748870

Elder, G. H., Jr., & Conger, R. D. (2000). *Children of the land: Adversity and success in rural America*. Chicago, IL: University of Chicago Press. http://dx.doi.org/10.7208/chicago/9780226224978.001.0001

Fritz, M. S., Kenny, D. A., & MacKinnon, D. P. (2016). The combined effects of measurement error and omitting confounders in the single-mediator model. *Multivariate Behavioral Research, 51,* 681–697. http://dx.doi.org/10.1080/00273171.2016.1224154

Judd, C. M., & Kenny, D. A. (1981a). *Estimating the effects of social interventions*. Cambridge, UK: Cambridge University Press.

Judd, C. M., & Kenny, D. A. (1981b). Process analysis: Estimating mediation in treatment evaluation. *Evaluation Review, 5,* 602–619. http://dx.doi.org/10.1177/0193841X8100500502

Judd, C. M., Westfall, J., & Kenny, D. A. (2017). Experiments with more than one random factor: Designs, analytic models, and statistical power. *Annual Review of Psychology, 68,* 601–625. http://dx.doi.org/10.1146/annurev-psych-122414-033702

Kashy, D. A., & Kenny, D. A. (2000). The analysis of data from dyads and groups. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (pp. 451–477). New York, NY: Cambridge University Press.

Kenny, D. A. (1979). *Correlation and causality*. New York, NY: Wiley-Interscience.

Kenny, D. A. (2020). *Interpersonal perception: The foundation of social relationships*. New York, NY: Guilford Press.

Kenny, D. A., & Cook, W. (1999). Partner effects in relationship research: Conceptual issues, analytic difficulties, and illustrations. *Personal Relationships, 6,* 433–448. http://dx.doi.org/10.1111/j.1475-6811.1999.tb00202.x

Kenny, D. A., & Judd, C. M. (1984). Estimating the nonlinear and interactive effects of latent variables. *Psychological Bulletin, 96,* 201–210. http://dx.doi.org/10.1037/0033-2909.96.1.201

Kenny, D. A., & Judd, C. M. (1986). Consequences of violating the independence assumption in analysis of variance. *Psychological Bulletin, 99,* 422–431. http://dx.doi.org/10.1037/0033-2909.99.3.422

Kenny, D. A., & Judd, C. M. (2014). Power anomalies in testing mediation. *Psychological Science, 25,* 334–339. http://dx.doi.org/10.1177/0956797613502676

Kenny, D. A., & Judd, C. M. (2019). The unappreciated heterogeneity of effect sizes: Implications for power, precision, planning of research, and replication. *Psychological Methods*. Advance online publication. http://dx.doi.org/10.1037/met0000209

Kenny, D. A., & La Voie, L. J. (1984). The social relations model. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 18, pp. 142–182). Orlando, FL: Academic Press.

Kenny, D. A., & Ledermann, T. (2010). Detecting, measuring, and testing dyadic patterns in the actor-partner interdependence model. *Journal of Family Psychology, 24,* 359–366. http://dx.doi.org/10.1037/a0019651

Kenny, D. A., & Milan, S. (2012). Identification: A non-technical discussion of a technical issue. In R. Hoyle, D. Kaplan, G. Marcoulides, & S. West (Eds.), *Handbook of structural equation modeling* (pp. 145–163). New York, NY: Guilford Press.

Kenny, D. A., & Zautra, A. (1995). The trait-state-error model for multiwave data. *Journal of Consulting and Clinical Psychology, 63,* 52–59. http://dx.doi.org/10.1037/0022-006X.63.1.52

Kenny, D. A., & Zautra, A. (2001). Trait-state models for longitudinal data. In A. Sayer & L. M. Collins (Eds.), *New methods for the analysis of change* (pp. 243–263). Washington, DC: American Psychological Association. http://dx.doi.org/10.1037/10409-008

Kim, Y., & Steiner, P. (2019). Gain scores revisited: A graphical models perspective. *Sociological Methods & Research*. Advance online publication. http://dx.doi.org/10.1177/0049124119826155

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Jr., Bahník, S., Bernstein, M. J., . . . Nosek, B. A. (2014). Data from investigating variation in replicability: A "many labs" replication project. *The Journal of Open Psychology Data, 2,* e4. http://dx.doi.org/10.5334/jopd.ad

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams, R. B., Jr., Alper, S., . . . Nosek, B. A. (2018). Many Labs 2: Investigating variation in replicability across sample and setting. *Advances in Methods and Practices in Psychological Science, 1,* 443–490. http://dx.doi.org/10.1177/2515245918810225

Klumb, P., Hoppmann, C., & Staats, M. (2006). Work hours affect spouse's cortisol secretion—For better and for worse. *Psychosomatic Medicine, 68,* 742–746. http://dx.doi.org/10.1097/01.psy.0000233231.55482.ff

Linden, A., & Hönekopp, J. (2019, March 13). Heterogeneity in the results of close and conceptual replications: Implications for scientific progress and practical applications. *ZPID (Leibniz Institute for Psychology Information)*. Advance online publication. http://dx.doi.org/10.23668/psycharchives.2398

Marsh, H. W. (1992). Self-Description Questionnaire (SDQ) II: A theoretical and empirical basis for the measurement of multiple dimensions of adolescent self-concept. Penrith, Australia: SELF Research Centre, University of Western Sydney.

Maxwell, S. E., & Cole, D. A. (2007). Bias in cross-sectional analyses of longitudinal mediation. *Psychological Methods, 12,* 23–44. http://dx.doi.org/10.1037/1082-989X.12.1.23

McShane, B. B., Tackett, J. L., Böckenholt, U., & Gelman, A. (2019). Large scale replication projects in contemporary psychological research. *The American Statistician, 73,* 99–105. http://dx.doi.org/10.1080/00031305.2018.1505655

Morse, G. A., Calsyn, R. J., Allen, G., & Kenny, D. A. (1994). Helping homeless mentally ill people: What variables mediate and moderate program effects? *American Journal of Community Psychology, 22,* 661–683. http://dx.doi.org/10.1007/BF02506898

Moshontz, H., Campbell, L., Ebersole, C. R., Chartier, C. R., IJzerman, H., Urry, H. L., . . . Chartier, C. R. (2018). The Psychological Science Accelerator: Advancing psychology through a distributed collaborative network. *Advances in Methods and Practices in Psychological Science, 1,* 501–515. http://dx.doi.org/10.1177/2515245918797607

Muthén, B., & Asparouhov, T. (2015). Causal effects in mediation modeling: An introduction with applications to latent variables. *Structural Equation Modeling, 22,* 12–23. http://dx.doi.org/10.1080/10705511.2014.935843

Pearl, J., & MacKenzie, D. (2018). *The book of why*. New York, NY: Basic Books.

Riggs, S. A., Cusimano, A. M., & Benson, K. M. (2011). Childhood emotional abuse and attachment processes in the dyadic adjustment of dating couples. *Journal of Counseling Psychology, 58,* 126–138. http://dx.doi.org/10.1037/a0021319

Rohrer, J. M. (2018). Thinking clearly about correlations and causation: Graphical causal models for observational data. *Advances in Methods and Practices in Psychological Science, 1,* 27–42. http://dx.doi.org/10.1177/2515245917745629

Rosenberg, M. (1965). *Society and the adolescent self-image*. Princeton, NJ: Princeton University Press. http://dx.doi.org/10.1515/9781400876136

Schuirmann, D. J. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics, 15,* 657–680. http://dx.doi.org/10.1007/BF01068419

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton Mifflin.

Simons, D. J., Holcombe, A. O., & Spellman, B. A. (2014). An introduction to registered replication reports at Perspectives on Psychological Science. *Perspectives on Psychological Science, 9,* 552–555. http://dx.doi.org/10.1177/1745691614543974

Wagner, J., Lüdtke, O., & Trautwein, U. (2016). Self-esteem is mostly stable across young adulthood: Evidence from latent STARTS models. *Journal of Personality, 84,* 523–535. http://dx.doi.org/10.1111/jopy.12178

Warner, R., Kenny, D. A., & Stoto, M. (1979). A new round robin analysis of variance for social interaction data. *Journal of Personality and Social Psychology, 37,* 1742–1757. http://dx.doi.org/10.1037/0022-3514.37.10.1742

Wong, V. C., Steiner, P. M., & Anglin, K. L. (2018). What can be learned from empirical evaluations of nonexperimental methods? *Evaluation Review, 42,* 147–175. http://dx.doi.org/10.1177/0193841X18776870

Received April 16, 2019
Revision received July 26, 2019
Accepted July 31, 2019 ■