In press March 18, 2019 – accepted draft, subject to final copy-editing
Advances in Methods and Practices in Psychological Science

Evaluating Effect Size in Psychological Research: Sense and Nonsense

David C. Funder and Daniel J. Ozer

University of California, Riverside

Author Notes

David C. Funder, Department of Psychology, University of California, Riverside; Daniel J. Ozer, Department of Psychology, University of California, Riverside

Preparation of this article was aided by National Science Foundation Grant BCS-1528131, David Funder, Principal Investigator. Any opinions, findings, and conclusions or recommendations expressed in this article are those of the individual researchers and do not necessarily reflect the views of the National Science Foundation.

DCF and DJO jointly generated the ideas for this article and jointly wrote the manuscript.

Correspondence concerning this article should be addressed to David C. Funder or Daniel J. Ozer, Department of Psychology, University of California, Riverside, CA 92521.
Emails: david.funder@ucr.edu,daniel.ozer@ucr.edu

Abstract

Effect sizes are underappreciated and often misinterpreted – the most common mistakes being to describe them in ways that are uninformative (e.g., using arbitrary standards) or misleading (e.g., squaring effect size *r*'s). We propose that effect sizes can be usefully evaluated in comparison with well-understood benchmarks or in terms of concrete consequences. In that light, we conclude that, when reliably estimated (a critical consideration), an effect of r =.05 is "very small" for the explanation of single events but potentially consequential in the not-very long run, r = .10 is still "small" at the level of single events but potentially more ultimately consequential; r = .20 is "medium" and of some use even in the short run and therefore even more important; and an effect size of r = .30 is "large" and potentially powerful in the short and long run. A "very large" effect size (r = .40 or greater) in the context of psychological research is likely to be a gross overestimate rarely found in a large sample or in a replication. Our goal is to help to move effect sizes from numbers that are ignored, reported without interpretation, or interpreted superficially or incorrectly, to aspects of research reports that can inform the application and theoretical development of psychological research.

Keywords: benchmarks, correlation, effect size, evaluation

Evaluating Effect Size in Psychological Research: Sense and Nonsense

Nonsense: Words or language having no meaning or conveying no intelligible ideas.
- Merriam-Webster

Psychological research has a long tradition of evaluating findings according to whether they are statistically "significant" or not; more recently, increasing attention has been paid to the size as opposed to the significance of an effect (e.g., Cumming, 2012). Effect size refers to the magnitude of the relation between the independent and dependent variables, and is separable from statistical significance in the sense that a highly significant finding could describe a small effect, and vice versa, depending on the N (sample size) of the study. Students are routinely taught how to calculate and interpret significance levels; they are less often taught how to calculate effect sizes or, even more rarely, how to evaluate them. This neglect of effect size persists into the research careers of many psychologists.

Much of the published literature reflects the continued neglect. Although many journals now require that effect sizes be reported, and researchers (usually) dutifully follow this requirement, they often ignore them otherwise. When researchers do draw implications from effect sizes, the interpretations they offer are, more often than not, superficial, uninformative, misleading, or completely wrong. In sum, effect sizes are widely unappreciated and often misunderstood, even by professional researchers.

Current research in psychological methods[1] has not been particularly helpful in this regard. Much of it concerns the development, specification, and testing of ever more elaborate or precise quantitative models, while only occasionally addressing the concerns of substantive researchers grappling with the "solved" problem of reliably detecting and measuring simple bivariate effects. When such problems do come to the fore, as in recent methodological articles and blog posts debating the threshold for rejecting the null hypothesis – should it be at $p < .05$ or $p < .005$ (Benjamin et al., 2017) – little reference is made to effect size. Despite exhortations to report effect sizes (e.g. Cumming, 2012), frank discussions of how to evaluate them remain surprisingly rare (but see Lakens, Scheel & Isager, 2018). The purpose of the present article is to help to remedy this imbalance.

Effect Size

The two most commonly used measures of effect size are Cohen's d and Pearson's *r*. The former, typically used to characterize the differences in means between experimental groups, is the mean

[1] E.g., as published in *Psychological Methods.*

difference divided by the pooled standard deviation. The latter, the correlation coefficient, is typically used to characterize the degree to which one variable can be predicted from another. These two measures of effect size can be algebraically converted from one to the other; for simplicity and consistency, the present article will focus on *r*.

Although for a considerable period of psychology's history it was common practice to report *p*-levels in the absence of an effect size, in recent years the reporting of effect size has become mandated by most journals. For example, the publication manual of the American Psychological Association, which covers many of the most visible outlets for psychological research, now states that reporting effect size is "almost always necessary" (2010, p. 34) and indeed, most (not all) articles in APA journals obediently report some measure, parenthetically, usually alongside the *p*-value. This mandated reporting is not always done enthusiastically. In a personal communication[2] with an author of this article, a prominent and widely published social psychologist wrote

> "…the key to our research… [is not] to accurately estimate effect size… When I am testing a theory about whether, say, positive mood reduces information processing in comparison with negative mood, I am worried about the direction of the effect, not the size. But if the results of such studies consistently produce a direction of effect where positive mood reduces processing in comparison with negative mood, I would not at all worry about whether the effect sizes are the same across studies or not, and I would not worry about the sheer size of the effects across studies. This is true in virtually all research settings in which I am engaged. I am not at all concerned about the effect size" (quoted in Funder, 2013).

This is not an unusual opinion; similar comments can be found in a number of articles and blog posts. However, it has two problems. First and most obviously, this researcher routinely uses *p*-levels to evaluate whether or not to be confident that his/her study has obtained a meaningful result. Given the study's N, to set a threshold *p*-level for accepting a result is exactly the same thing as setting a minimum effect size for the same decision. For example, in a two-group experimental study with 60 subjects, setting a two-tailed *p*-level threshold of .05 is equivalent to setting an effect size threshold of $r = .304$. Only slightly less obviously, the social psychological literature is filled with (usually non-numerical) references to effect size, such as claims that certain manipulations can have "large" or even "surprisingly large" effects, or that (in the case of the so-called Fundamental Attribution

[2] The writer of this communication gave permission to quote it, but not to identify the writer by name.

Error) most people believe personality traits have "larger" effects than they really do. Such claims proceed in an empirical vacuum without some kind of quantitative measure of effect size.[3]

The Two Most Common Ways to Interpret Effect Size

When effect sizes are interpreted, the interpretation traditionally proceeds in one of two ways. The first of these is literally nonsensical (in the meaning expressed in the aphorism at the top of this article), and the other is seriously misleading.

Cohen's Standards. The nonsensical but widely used interpretation of effect size is the famous standard set by Cohen (1977, 1988), who (in terms of *r*) set .10 as a "small" effect, .30 as a "medium" effect and .50 as a "large" effect. Cohen reluctantly used these conventions in the context of power analysis "only when no better basis…is available" (p. 25), and later told friends he actually regretted having suggested them at all (R. Rosenthal, personal communication). He had good reason for this regret. The terms "small," "medium" and "large" are meaningless in the absence of a frame of reference. They immediately require answers to at least one of two questions: (1) small, medium or large, compared to what? Or (2) small, medium or large, for what purpose? (We shall return to these questions later in this article.)

Squaring the Correlation. As bad as these decontextualized criteria are, the other widely used way to evaluate effect size is arguably even worse. This method is to take the reported *r*, and square it. For example, an *r* of .30, squared, yields the number .09 as the "percentage of variance explained," and this conversion, when reported, often includes the word "only," as in "the .30 correlation explained only 9% of the variance."

We suggest that this calculation has become widespread for three reasons. First, it is easy arithmetic that gives the illusion of adding information to a statistic. Second, the common terminology of "variance explained" makes the number sound as if it does precisely what one would want to it do, with the word "explained" evoking a particularly virtuous response. Third, the context in which this calculation is often deployed allows writers to disparage certain findings that they find incompatible with their own theoretical predilections. For one prominent example, Walter Mischel's classic critique of personality psychology complained that the "personality coefficient" of .30, described by him as the upper limit for

[3] Indeed, in the case of the Fundamental Attribution Error, despite the fame and influence of this idea, we are unaware of a single in study in which lay beliefs about the influence of personality were compared quantitatively with its actual influence.

the predictability of behavior from trait measurements[4], "accounts for less than 10 percent of the relevant variance" (Mischel, 1968, p. 38). As Abelson (1985) observed, "it is usually an effective criticism when one can highlight the explanatory weakness of an investigator's pet variables in percentage terms" (p. 129).

The variance "explained" by the squared *r* is about squared deviations of the variable from its mean. The squaring of the *r* changes the scale of the effect from the original units to squared units. One can search statistics textbook after textbook without finding any attempt to explain (as opposed to assert) why these squared units are appropriate for evaluating effect size (i.e., why one would want to account for variance rather than standard deviation). The squared correlation may have some utility as a measure of model fit but the original, unsquared *r* reflects the size of the effect on the metric of the original measured units.

Consider the difference between nickels and dimes. An example introduced by Darlington (1990) shows how their difference in value can be distorted by traditional analyses. Imagine a coin-tossing game in which one flips a nickel and then a dime, and receives a 5-cent or 10-cent payoff (respectively) if the coin comes up heads. From the payoff matrix illustrated below (in which 1 denotes heads and 0 denotes tails), a correlation can be calculated between the nickel column and the payoff column (r = .4472), and between the dime column and the payoff column (r = .8944). If one squares these correlations to calculate the traditional "percentage of variance explained," the result is that nickels explain exactly 20% of the variance in payoff, and dimes explain 80%. And indeed, these two numbers do sum neatly to 1.0, which helps to explain their attractiveness in certain analytic contexts. But if they lead to the conclusion that dimes matter four times as much as nickels, the numbers have obviously been misleading. The two *r*'s afford a more informative comparison; .8944 is exactly twice as much as .4472. This is the sense in which a correlation of .4 reveals an effect twice as large as a correlation of .2; moreover, half of a perfect association is .5, not .707 (Ozer, 1985, 2007). Squaring the *r* is not merely uninformative; for purposes of evaluating effect size, the practice is actively misleading.

Nickel Tossed	Dime Tossed	Total Payoff
1	**1**	**.15**
1	**0**	**.05**
0	**1**	**.10**
0	**0**	**0**

[4] This was later raised to .40 by Nisbett (1980).

Towards Useful Interpretations of Effect Size

How can effect sizes be interpreted in a way that adds or provides meaning? We suggest two answers. The first is to use a benchmark, and the second is to estimate consequences.

Benchmarks

The idea of using benchmarks to evaluate effect size is that the magnitude of a finding can be illuminated by comparing it to some other finding that is already well-understood (or that at least is widely believed to be well-understood). All of the benchmarking strategies summarized below have the same aim: to help a reader attain an intuitive "feel" for the meaning of an effect size. In the same way we immediately gauge whether somebody is tall or short by comparing the person to the other people we know, with a knowledge of the sizes of classic findings, average findings, or other effects that are understood through everyday experience, we can approach a realistic appreciation of the meaning of a particular research result. Cohen (1988) used this strategy to justify his labeling numerical effect size values in terms of small, medium and large; likening a small effect to several specific effects, such as the mean height difference between 16 and 17 year-old girls. Medium effects were characterized as those "visible to the naked eye (p. 26)," though it seems he may have grossly overestimated the sensitivity of observers to at least some characteristics (Ozer, 1993). Large effects were vivified with a comparison to the difference between the mean IQ of college graduates with those with just a 50-50 chance of graduating from high school. One might well quibble with the choices of examples offered by Cohen (1988), but given the context (few others were talking about effect size), it would seem more fruitful to consider other benchmarking approaches.

Classic Studies. One example of the benchmark approach was an analysis reported by the present authors some years ago (Funder & Ozer, 1983). This article performed a simple re-analysis of three classics of the psychological literature: the finding of a reverse incentive effect on attitude change, by Festinger and Carlsmith (1959), the studies of bystander intervention by Darley & Latané (1968) and Darley and Batson (1967), and the demonstrations of experimentally-induced obedience by Milgram (1975). In each case, from the reported findings we simply computed an effect size *r* that reflected the degree to which the dependent variable (attitude change, helping, or obedience, respectively) was affected by the manipulated independent variable (incentive, hurry and number of bystanders, and distance of experimenter and victim, respectively). In each case, the resulting effect size *r* fell between .36 and .42.

This result should not have been surprising but it was, in a zeitgeist when a common complaint about personality traits was that their correlations with behavioral outcomes seldom exceeded .40 (e.g., Nisbett,

1980). And some writers at the time misinterpreted the implication of our calculations, in our view, by concluding that the calculations implied that these studies also "only" found "small" effects after all (or more disastrously, "situations aren't important either"). Our own view was that these studies were and remain classics of the social psychological literature and nobody, certainly nobody at the time, doubted that the effects they reported were foundation stones of social psychology that should be taught to every student of the topic. We simply thought it was worth knowing that their reported effect sizes were in roughly the same range as the purported ceiling for effects of personality.

Other Well-established Psychological Findings. A later, similar, but much broader set of re-analyses also calculated the effect size *r*'s of well-established findings in psychology (Richard et al., 2003). To name a few examples, scarcity increases the perceived value of a commodity (r = .12), people attribute failures to bad luck (r = .10), communicators perceived as more credible are more persuasive (r = .10), and people in a bad mood are more aggressive (r = .41). A reader is free to decide whether or not to interpret any of these findings as reliable or important, but to the extent that one does, then the associated effect sizes provide a useful benchmark for interpreting other findings that one encounters in the literature.

A similar analysis was performed by Roberts et al. (2007), who compared the validity of personality traits for predicting mortality, divorce and occupational success with the well-established predictors social-economic status (SES) and intelligence (IQ). The result was that the "magnitude of the effects…was indistinguishable" (p. 313). Even more striking, perhaps, was that, for the prediction of mortality, the estimated *r*'s for these effects – of both kinds – ranged no higher than r = .24 and for the most part fell below r = .10.

Comparisons with "All" Studies. Even broader efforts have sought to find effect size benchmarks from the average of comprehensive reviews of social and personality psychology. The ambitious effort by Richard et al., cited above, also calculated an average effect size of *all* the published effects in the social psychological literature that the authors were able to survey, and the result was r = .21. A parallel but less extensive project surveyed the personality literature and came up with precisely the same average effect size: r = .21 (Fraley & Marks, 2007). Of course, both of these results are very likely to be overestimates of the true effects of the variables studied, because of publication bias that privileges significant (and so on average larger) findings to appear in the literature. Therefore, a researcher who completes a new study that obtains an r = .21 can be fairly confident that this is a larger effect than typically found.

A more recent and very large project reviewed 708 meta-analytically derived correlations from the literatures of both social and personality psychology, and found that the average effect size *r* was .19, and

that *r*'s of .11 and .29 fell into the 25th and 75th percentiles, respectively (Gignac & Szodorai, 2016). The authors suggested that, in this light, the Cohen guidelines could be recast as .10, .20, and .30 being "small, typical, or relatively large," correlations, respectively (p. 74).

Comparisons with Intuitively Understood non-Psychological Relations. If from ordinary life experience or broader reading one has developed a sense of how strongly one variable is related to another, then this understanding can also be used as an aid to the intuitive appreciation of a research finding. For example, do you take antihistamines to combat runny nose and sneezing? How well do they work? According to one estimate, the effect size of the relationship between antihistamine use and relief from these symptoms is equivalent to $r = .11$. Do you take a pain reliever to alleviate headaches? The relieving effect of nonsteroidal anti-inflammatory drugs (such as ibuprofen) on pain is not much different than the effectiveness of antihistamines on sneezing; $r = .14$. Other familiar benchmarks for intuitively calibrating effects include the tendency of men to weigh more than women, overall ($r = .26$), the tendency of places at high elevations to have lower average annual temperatures ($r = -.34$), and the correlation between height and weight for US adults ($r = .44$). And, for a really big one: the effect size of the average height difference between men and women is equivalent to $r = .67$ (all these findings are summarized by Meyer, et al., 2001, pp. 131-132).

Consequences

The Binomial Effect Size Display. Beyond comparisons to benchmarks, a more direct way to evaluate an effect size is in terms of its consequences, which in some cases can be numerically calculated. Perhaps the best-known and easiest to use of these methods is the Binominal Effect Size Display (BESD) introduced by Rosenthal and Rubin (1982). The BESD is used to illustrate the size of an effect, reported in terms of *r*, on a 2 x 2 table of outcomes. In its usual application, the process begins with assuming that a sample of 200 individuals has been divided into equal-sized two groups, one of which has experienced an intervention (such as a drug for a disease all 200 have) and one has not. It is further assumed, for the sake of illustration, that for half the individuals the intervention was successful and for the other half it was not. If the intervention (or drug) had no effect at all (effect size $r = 0$), the 2 x 2 table would look like this:

$r = 0$

	Successful outcome	Unsuccessful outcome	Total
Intervention	50	50	100
No intervention	50	50	100
Total	100	100	200

In Rosenthal and Rubin's favorite (hypothetical) example, the intervention comprises giving a drug or not, and the outcomes comprise being alive or dead at the end of the study. Less dramatically, but also perhaps more generally applicable, any pairing of a dichotomous predictor and dichotomous outcome can be analyzed in this way. The effect size *r* can easily be reflected on this table by multiplying it by 100 (to remove the decimal), dividing it by 2, adding 50, and placing the result into the upper left-hand corner. The remaining cells are determined by subtraction (since this table has 1 degree of freedom). 30 divided by 2 is 15, plus 50 is 65 and so if $r = .30$, the table looks like this:

$r = .30$

	Successful outcome	Unsuccessful outcome	Total
Intervention	65	35	100
No intervention	35	65	100
Total	100	100	200

Some readers, traditionally trained to think of .30 correlations as "explaining only 9% of the variance" might be surprised to learn that an effect of this size will yield almost twice as many correct predictions – or live outcomes – as incorrect ones. More specifically, a table such as this, when combined with cost data for interventions and outcomes, could be used to calculate the utility of an intervention or of a predictive instrument in concrete, monetary terms. It could also be used, as in Rosenthal and Rubin's own example, to assess the number of lives that could be saved by a health intervention. A later analysis calculated that the correlation of r = .03 between taking aspirin after a heart attack and prevention of future heart attacks implied the prevention of 85 attacks in a sample of 10,845 individuals (Rosenthal, 1990). Again, less dramatically, it could be used to calculate the payoff from using an ability or personality test to select employees. In a similar manner, the Taylor-Russell tables (Taylor & Russell, 1939) have long been used by industrial psychologists to combine the validity of a selection instrument

with the selection ratio (the proportion of applicants hired) to predict the percentage of hired employees who will be successful on the job.[5]

Consequences in the Long Run. In a classic analysis (which is nonetheless not as widely known as it should be), subtitled "When a Little is a Lot," the well-known cognitive psychologist Robert Abelson calculated the correlation between a single at-bat for a major league baseball player, and his overall batting average. The result was equivalent to $r = .056$[6]. He was so surprised by this result that he exclaimed (in print) "What's going on here?" (Abelson, 1985, p. 131). It is testimony to the degree to which the variance explanation ritual had become mindlessly entrenched even in the thinking of sophisticated researchers that he confessed that his "first reaction to this result [was] incredulity… My personal intuition was jarred by this result, which seems much too small" (p. 131). The mystery appeared to deepen when he observed that almost all major league baseball players have season averages within a limited range, between about .200 and .300.

However, the resolution to what Abelson characterized as a "paradox" (p. 131) turned out to be rather simple. The typical major league baseball player sees about 550 at-bats in a season. The consequences cumulate. This cumulation is enough, it seems, to drive the outcome that a team staffed with .300 players is likely on the way to the playoffs, and one staffed with .200 players is at risk of coming in last place. The salary difference between a .200 batter and a .300 batter is in the millions of dollars for good reason.

For another example, a large study that tracked 2 million financial transactions across more than 2000 people found that the correlation between an individual's extraversion score and the amount he or she spent on holiday shopping is $r = .09$ (Weston et al., 2018). While this fact might not be very consequential for a single individual, multiply the effect identified with this correlation by the number of people in a department store the week before Christmas, and it becomes obvious why merchandisers should care deeply about the personalities of their customers.

The overall implication, as Abelson noted, is that seemingly small effects can matter "in the long run, albeit not very consequentially in the single episode" (1985, p. 133). In particular, a psychological process

[5] A comparison of the BESD to the Taylor-Russell tables will show some discrepancies even in the case of equal marginal proportions, due to the median splits of continuous distributions imposed but not accounted for by the BESD as they are by the Taylor- Russell tables. In the BESD, the values of the main diagonal of the contingency table can be computed by 50+ 100*r*/2, then a close approximation to the Taylor- Russell tables can be obtained by using 50 + 100r/3.

[6] Actually, he reported that the "percentage of variance in any single batting performance explained by batting skill" is .00317; the .056 figure is the square root of that number.

that affects the behavior of a single individual repeatedly[7] over time, or, analogously, the behavior of many individuals simultaneously on a single occasion, can have hugely important implications.

Relevance for Psychological Research

Abelson's illustration of how seemingly small effects can cumulate has important implications for psychology. Every social encounter, behavior, reaction and feeling a person has, could be considered a psychological "at-bat." And imagine how many of those occur in a day, a week, a year, or a lifetime – certainly many more than the 550 or so a ball player gets in a year. Any psychological variable that affects any of these, every time it happens, will have an effect that could cumulate over time with important consequences for numerous life outcomes including (to name just a few examples) popularity and social success, physical health, financial success, personal relationships, and one's overall quality of life[8].

Individual Difference Research

The relevance of the cumulation of small effects over time is particularly obvious for research on individual differences, such as abilities or personality traits. If a stable trait – such as extraversion, agreeableness, or conscientiousness, for example – affects much of what you do even in a small way, its consequences can add up very and perhaps surprisingly quickly. Analyses of the effects of personality on life outcomes have focused on long-term consequences such as health, relationship success, quality of life and – that ultimate long-term consequence – longevity (Friedman et al., 1993; Ozer & Benet-Martínez, 2006; Roberts et al., 2007). But Abelson's analysis suggests that one might need much less than a lifetime for noticeable consequences of stable personality traits to appear. A correlation of .05 translates to large consequences with 550 at-bats; how long does it take for a person to experience, for example, 550 interpersonal encounters?

Consider a student moving away from home to college, and meeting the fellow residents of her dormitory for the first time. Assume she is highly agreeable. How long will it take before she finds herself enjoying the enhanced popularity that is the reliable long-term result of this trait (Ozer & Benet-Martínez, 2006)? A back-of-the envelope calculation suggests that if the correlation between agreeableness and an individually successful social interaction is r = .05 (which is a hypothetical, conservative estimate[9]), and

[7] With consequences that cumulate, a point we will consider further below.

[8] Of course, a psychological event with a large effect size could be important even if it only occurs once; such events – such as a traumatic experience – may be rare but powerful.

[9] In a recently gathered international data set with an N=15,432, the correlation between Agreeableness and experiencing a single situation as "enjoyable," and "arousing positive emotions" is r = .07 for each outcome, and

if she has 20 social interactions a day, then the consequences for her popularity will be as noticeable as the consequences of batting ability for a baseball player's success at the end of the season, in less than a month[10].

Even more remarkably, Epstein (1979) demonstrated how broad outcome criteria could be predicted with surprising precision from broad, aggregated predictor variables. For example, he showed that a person's average behavior over a period of 14 days could be predicted with a correlation equivalent to r = .80 to r = .90, when the predictor was the person's average behavior over a preceding period of 14 days (Epstein, 1979, p. 1123). The moral of his demonstration was that an appropriate and realistic target for behavioral prediction is not what a person does on one day or in one situation, but what he or she does in the not-very long run.

Experimental Research

The relevance of the way effects can cumulate over time is perhaps less obvious for experimental research but it is fundamentally no different. If a psychological process is experimentally demonstrated, and this process is found to appear reliably, then its influence could in many cases be expected to accumulate into important implications over time or across people even if its effect size is seemingly small in any particular instance.

For example, a factor that has a small influence on the degree to which a person can accomplish self-control every time he or she experiences fatigue – which is perhaps not every day, but certainly not rare – will become a psychologically important fact to understand about what goes on when people are tired[11] (see further discussion of this point, below). Or consider, for another example, the recent meta-analytic conclusion that effect of a growth mindset intervention on student achievement is r = .08, a size the reviewers described as "weak" (Sisk et al. 2018, p. 549). But this effect can also be calculated to imply an average increase of GPA (on the traditional 4-point scale) of .1 of a point, which when aggregated across all the students in a class, a school, or a school district could translate to a lot of increased student achievement (Dweck, 2018, Gelman, 2018)[12]. Or, for a final example, an aspect of a communication that

---

the correlation with experiencing the situation as "anxiety-inducing" or "hostile" is r = -.08 for each outcome (International Situations Project, 2018).

[10] To be exact, 550/20= 27.5 days.

[11] This is the phenomenon sometimes called "ego-depletion," which in a large set of replication studies was reported to yield an average effect size equivalent to r = .05, almost exactly the same as Abelson's baseball example (Vohs, 2018).

[12] Gelman (2018) further points out that this effect size could imply a change of 1 full GPA point for 10% of the students in the sample, and no change at all for the others. An effect that is small on average could still have large effects for particular individuals.

(reliably) makes it even a tiny bit more persuasive may become important when it is conveyed to millions of people. Imagine, for example, that a political consultant is purchasing time for a TV ad that will be seen by 30 million people, and is choosing between two possibilities that experimental research has shown differ in their effectiveness with an effect size *r* of .05. The choice is obviously consequential. This is the sense in which experimentally demonstrated phenomena could cumulate in their importance even if their one-time (or one-person) effect sizes are in the range traditionally dismissed as weak.

A long-standing tradition in experimental social psychology has been to try to recreate real-world situations in the laboratory (Aronson & Carlsmith, 1968). Influential studies have simulated circumstances in which a person appears to be in distress in order to assess the conditions under which a bystander might intervene, a person is given dire orders to harm another person in order the assess the conditions under which obedience or disobedience becomes more likely, or a person is given an initial (false) impression of someone he or she is about to meet, in order to assess the conditions under which this impression becomes self-fulfilling. Such research has become increasingly rare in recent years, perhaps because it is difficult to do for operational and ethical reasons, and also because easier methods of research such as gathering responses to computer-presented stimuli have become widely available (Baumeister, Vohs & Funder, 2007).

Indeed, to capture a meaningful aspect of social experience in a psychological laboratory, for even a few minutes, is a remarkably ambitious and even daunting goal. Some experimental research of this sort turns out to fail to replicate, and the findings not to be reliable after all. But in the cases where some aspect of a situation does turn out to affect behavior, and the finding is reliable across experimental attempts and different laboratories, then lightening has been caught in a bottle[13], and it is not wise nor even realistic to demand a "large" effect size (Gelman, 2018). Under the circumstances, to find anything at all can be impressive (Prentice & Miller, 1992).

When Effects Do (and Do Not) Cumulate

The foregoing discussion applies to circumstances in which the effects measured by a research study can be expected to cumulate over time, situations or individuals. Small effects accumulate into large ones in at least some, probably many, but certainly not all circumstances. This cumulation can occur across time and occasions for a given individual, and across individuals at a single time or occasion.

[13] "Capturing something powerful and elusive and then being able to hold it and show it to the world" (Urban dictionary, https://www.urbandictionary.com/define.php?term=lightning%20in%20a%20bottle)

An unambiguous example of cumulation across time and situations for an individual was the case, considered earlier, of the baseball batting average, in which hits add up (in the not-very long run) into runs, and runs (also in the not-very long run) add up into won games. Another example of cumulation, that seems almost as clear to us, was the way the (even slightly) larger probability of a friendly act by an agreeable person can lead, before, too long, to an enhanced social reputation. More generally, precisely because they are consistent over time and across situations, the influences of personality on behavior can confidently be expected to affect consequential social, occupational, and health outcomes, and in fact they do (Ozer & Benet-Martínez, 2006).

Not all cases are as clear-cut as these, however. It is not difficult to think of examples in which repeated effects fail to increase, increase non-linearly, or even reverse in their consequences over time and occasions. The well-known Weber-Fechner and Yerkes-Dodson principles describe how responses to increases in level of a stimulus or motivation tend to level off or even reverse; the principle of habituation posits that responses to a repeated stimulus will eventually cease altogether. Cognitive systems of emotional regulation and physiological systems enforcing homeostasis, similarly, can reduce or eliminate the effect of repeated stimuli. Another potential complication to cumulation is the Matthew effect, which posits that the accumulation of advantages (or other consequences) from a psychological process can actually accelerate, perhaps differently over time for different individuals.

Even in cases such as these, however, in which the strength of the effect itself does not build steadily over time, the *consequences* of the underlying process still might. In the case of "ego depletion," for example, imagine a person who dislikes her job so much that she comes home every evening in a state of psychological fatigue that makes her more likely (with an effect size equivalent to $r = .05$) to have a short fuse in stressful conversations with her spouse. Even if recovery processes repair the deficit in self-control by the next morning (e.g., Inzlicht & Schmeichel, 2012), the even-slightly increased daily probability of marital friction would seem likely to have important consequences in the medium-run (say, about two years, or nearly 550 work days). And, on a theoretical level, an underlying process that can affect interpersonal interactions with a small but real probability on each individual occasion, can be important for understanding relationships and many other outcomes that are the long-term result of many interactions.

As one thinks through examples such as this – and whether or not a reader agrees with this or any other particular interpretation – a common gap in psychological theorizing becomes evident: When and in what ways can individual differences, situational variables, and their underlying processes – which may have small effects on single occasions – be expected to cumulate in their strength and/or consequences? And

what sorts of processes will *not* cumulate in their strength or consequences? When and if psychological theorizing begins to take more careful account of effect sizes – one hoped-for goal of this article – then attention to these questions will become critical. The metaphor of the psychological "at-bat" may apply in many cases, but surely not all, and theories could be more helpful than they currently are, in identifying them.

Reliably Estimating Effect Sizes

The foregoing analysis is also based on a presumption that the effect size in question is, in fact, reliably estimated. This is a big presumption, and a critical concern when the effect size is in the range traditionally regarded as "small." While the difference between an *r* of .30 and .40 might not be terribly important for most theoretical or practical purposes, the difference between an *r* of .00 and .10 surely is. In that light, it is sobering to observe that the 95% confidence interval of *r* =.10 will not quite exclude a value .00 with an N of 400, and to exclude .00 from the 95% confidence interval of *r* =.05 requires an N of 1500. Fortunately, there are other ways to establish effect sizes besides single studies with very large N's. Meta-analytically, a series of diverse studies of a topic that all get effect sizes within a narrow range (and in the same direction), even if the average effect is "small," can provide some reasonable degree of confidence that the effect has been usefully estimated. In any event, it is clear that the precision of the estimate of the effect size becomes more important the smaller the effect size is.

Other, non-statistical considerations can be of concern as well. Smaller effects are more at risk of being the sole product of an artifact rather than the process under investigation. For example, experimenter expectancy effects (Rosenthal, 2009), even if less powerful and more subtle than initially reported (Jussim, 2017), might be enough to account for effects in the range, for example, of the .08 effect of mindfulness interventions mentioned earlier in this article. This example illustrates another sense in which the precise estimate of smaller effects becomes crucial; not only are larger N's and more studies desirable, but care in eliminating potential confounding variables also becomes critically important. Other practices to reduce bias in analysis and reporting of research findings, such as preregistration of studies or registered reports, can also be helpful, because the importance of potential bias becomes larger when effects or N's are smaller.

Implications for Interpreting Research Findings

The foregoing has three important implications for how research findings should be interpreted.

We Should Not Automatically Dismiss "Small" Effects

One reason why experimental social psychologists, in particular, have seemed reluctant to report or to emphasize effect sizes might be that, because of their traditional training (which often includes squaring correlations to yield percentage of variance explained), they are taken aback by how small they seem. If readers of the psychological literature better understood the implications of effect size, apologies for the effect size one is reporting may no longer be necessary. Rather than an incentive structure that rewards performing selective analyses (*p*-hacking) in order to increase small effect sizes across the threshold of statistical significance, incentives would instead reward studies with large N's that unapologetically find and report "small" effect sizes with better precision and reliability – which is no small accomplishment, as explained above.

Indeed, one objection that is sometimes voiced to recommendations to gather large samples is that small, unimportant effects will become significant. We believe this objection is mistaken, because it is smaller effect sizes that (realistically) will turn out to be the ones that are more likely to have been correctly estimated, and other things being equal larger sample sizes are likely to provide more precise estimates regardless of the size of the effect.

Effect sizes will become more prominently and less reluctantly reported in experimental research, we believe, when researchers stop feeling (or being made to feel) defensive about them, and when explicit (rather than ritualized) discussions of the theoretical and practical implications of obtained effect sizes, of any magnitude, become more common. As reports of research begin to accumulate in the literature, with effect size reported in abstracts and perhaps even titles of articles, readers will begin to develop their own, experientially based and more realistic intuitions about what "small" and "large" really mean in this context.

We Should Be More Skeptical of "Large" Effects

As the flip side of the comment just made, the traditional neglect of effect size reporting has also allowed some implausibly large effects to sneak in under the radar. To name just one famous example, the effect of unscrambling words, referring to stereotypes of the elderly, on slowing down walking speed was found in two studies (with an N of 30 in each) to be equivalent to r = .48 and r =.38 (Bargh, Chen & Burrows, 1998). These numbers were not reported in the original article but are easily calculated from the reported *t*-statistics. If they had been reported, and appropriately evaluated in terms of benchmarks, questions

might have been raised about the plausibility of this effect really being about as large as the correlation between height and weight in US adults (as cited earlier, r = .44; Meyer et al., 2001)[14].

Small N studies have often reported anomalously large effect sizes. This might have been a sign, if heeded, that their overall reliability was not to be trusted. Because the confidence intervals of effect sizes in small studies are very wide, such studies can therefore be expected to sometimes produce large apparent effects that on replication turn out to be drastic overestimates (Cumming, 2012). A recent major project found that even for studies published in highly prestigious journals whose findings could be successfully replicated, the replication effect sizes were about half the size of the originals (Camerer et al., 2018). In our view, enough experience has already accumulated to make one suspect that small effect sizes, from large N studies, are the most likely to reflect the true state of nature.

We Should Be More Realistic about the Aim of any Program of Psychological Research

Looking across a room full of research psychologists at a professional meeting, it is possible to be struck by the thought that everyone there believes, usually with some justification, that what they are studying is important. As a result, every psychologist is prone to expect that the variable he or she is studying should have a large effect on cognition, emotion, or behavior. This is perhaps sometimes the case, but every researcher should also be aware that the psychologist sitting next to him or her may be studying a very different topic, but with the same expectation. And let's face the fact: human psychology is inherently complex, and there is only so much variation – in cognition, emotion, or behavior – to go around (Ahadi & Diener, 1989, De Boeck & Jeon, 2018).

How realistic is it to expect that any one research program, on any one topic or addressing any one psychological process, however real it might be, determines more than a small piece of what is really going on in the psychological world at large? Perhaps we should all lower our expectations a little (or a lot). As psychologists, we are in the business of predicting the results of experiential or behavioral at-bats, and should not be surprised or begrudge that we must share our predictive validity with other correlates and causes.

[14] Subsequent attempts to replicate this finding have been largely unsuccessful (e.g., Doyen, Klein, Pichon & Cleeremans, 2012).

Recommendations for Research Practice

Report Effect Sizes, Always and Prominently

Every study should report the effect sizes prominently. This is routine in individual differences research, where the Pearson *r* is ubiquitous, but even these articles could more strongly emphasize the actual size beyond the existence of the relationships they report. Experimental research has farther to go; it needs to move the effect sizes that are mandated to be reported out of reluctant, parenthetical mentions buried in the Results section, to the Abstract and Discussion sections as well. Over time, a base of experience will accumulate as readers of the literature – researchers and students alike – become gradually familiar with the effect sizes that are actually found in well-conducted research. A corollary of this recommendation is that the N of each study should be sufficient to have the effect size estimate be at least somewhat reliable.

A recent example is a meta-analysis of 761 effect sizes, gathered on a total N = 420,595 (Allen & Walter, 2018). The article reported – in its abstract – several relationships between personality traits and sexual behavior, including (among others) the correlation between extraversion and frequency of sexual activity (r = .17), agreeableness and sexually aggressive behavior (r = -.20) and conscientiousness and sexual infidelity (r = -.17). This is exactly the kind of reporting that illuminates not only the specific findings summarized, but helps to build a larger understanding of how big important effects can really be expected to be.

Do Studies with Larger N's (when possible)

As we have noted, an often-neglected complication in interpreting effect sizes is that the confidence interval of *r* is very wide with small samples. Schönbrodt and Perugini (2013) ran a series of Monte Carlo simulations that led them to conclude that "in typical scenarios sample size should approach 250 for stable estimates."

We believe that the effect size is information that should be reported and evaluated regardless of the N of the study. But the confidence interval should be reported as well, allowing evaluation to be informed by the necessary degree of uncertainty when N's are small. The ideal solution is to run studies with large N's. This is not always feasible with certain kinds of research or participant populations (Finkel, Eastwick & Reis, 2015). But an important priority should be to make studies as large as resources allow, and perhaps it would be wise to reallocate resources from numerous smaller studies to fewer larger ones. A few studies with larger N's are likely to produce more accurate and less confusing findings than will many studies with smaller N's. In particular, the recent history of social psychology illustrates the

bewildering welter of seemingly contradictory results that can emerge from a literature dominated by small-N studies.

Report Effect Sizes in Terms that are Meaningful in Context

The Pearson *r*, emphasized in the present article, is a standardized measure of effect size which means it has no reference to, and provides no information about, the units of measurement used in the study. An insufficiently recognized property of standardized measures of effect size like *r* is that they confound the consistency of an effect with the size of the effect. Imagine predicting annual salary from years of education in a heterogeneous sample of adults. It is possible, however fanciful in this example, that the correlation between years of education and income is nearly 1.0 yet a year of education might only be worth a dollar in annual income: All cases fall on a nearly flat regression line. The linear model fits very well, hence the large correlation; but the effect is very small, as would be shown by the raw regression coefficient. Or the discrepancy between model fit and effect size could be seen in the opposite fashion: On average a very large (steep slope) but highly variable (large standard error of estimate) effect of education on income. Fit of the linear model, or the consistency of the effect; and the slope of the regression line, or the size of the effect, are inherently confounded in standardized measures of effect size.

We are not the first to make the point (see, for example Cohen *et al*, 1999) that more meaningful measures of our variables would lead to more meaningful measures of their effects. The need to employ standardized measures of effect size arises from the use of arbitrary and intrinsically meaningless measurement units. We would be well served to be explicit about these units, and utilize raw effect size measures like mean differences or raw regression coefficients alongside our standardized measures of effect size, when possible. This would remind us of the ambiguities inherent with the standardized effect measures, and contribute toward the development of an interpretive framework for our most frequently used measurement units (Pek & Flora, 2017). With experience, even the meaning of a unit on a 7-point Likert scale might eventually become clear.

Moreover, there are cases, especially in applied research, when the unit of measurement does have an intrinsic meaning. For example, the mean differences in a countable health outcome, such as heart attacks, are meaningful in their own right and should be reported in preference to standardized measures such as correlations or relative risks. As an illustration, the Harding Center for Risk Literacy reports “fact boxes” that describe costs and benefits of health interventions in terms of concrete numbers such as the number of people who would benefit or be harmed by a screening or a drug (Harding Center, 2018a). One of their examples translates medical effect size statistics in the following manner: Consider a sample of 200

people suffering from acute bronchitis. If 100 of these people are given no treatment or a placebo, after 14 days 51 of them will still have a cough and 19 will feel ill in other ways (e.g., nausea). If the other 100 are given an antibiotic, 14 days later only 32 of them will still have a cough but 23 will feel ill otherwise (Harding Center, 2018b). This kind of format for presenting research results translates effect sizes into consequences people care and make decisions about.

Stop Using Empty Terminology

It is far past time for psychologists to stop squaring *r*'s in order to belittle the seemingly small "percentage of variance explained," and to stop mindlessly using the Cohen guidelines that even Cohen came to disavow. Ideally, words like "small" and "large" would be expunged from the vocabulary of effect sizes entirely, since they attach a subjective and often arbitrary label that adds no information to a number that can be quantitatively reported. This goal is probably unrealistic; indeed, in the present article we have unable to avoid the liberal use of these descriptive adjectives ourselves. But at the very least, it would be good to become in the habit of responding to characterizations of effect size as being small or large with questions such as, compared to what? Compared to what's usually found, to what other studies have shown, to what's useful to know, or another standard altogether? Whatever the standard of evaluation is, there ought to be one.

Revise the Cohen Guidelines

This is our most presumptuous recommendation, and we offer it somewhat tongue-in-cheek, but not entirely. It is abundantly clear that the traditional Cohen guidelines are much too stringent. And like Cohen, we think decontextualized guidelines are only appropriate for the most approximate of uses. But new guidelines can be proffered in the light of Abelson's demonstration of the not-so-long-term consequences of an effect size of $r = .05$, the illustration of the BESD of how a correlation in the range of $r = .30$ can almost double predictive validity beyond chance, the average sizes of effects in the published literature of social and personality psychology, and the sizes of other relationships between variables encountered in daily experience, such as the effectiveness of antihistamines or the association between height and weight.

We offer, therefore, the following New Guidelines: assuming the estimates are reliable (a critical concern, as already discussed), an effect with the size of $r = .05$ is "very small" for the explanation of single events but potentially consequential in the not-very long run, $r = .10$ is still "small" at the level of single events but potentially more ultimately consequential; an effect size of $r = .20$ is "medium" and of some use even in the short run and therefore even more important; and an effect size of $r = .30$ is "large" and potentially

powerful in the short and long run[15]. A "very large" effect size (r = .40 or greater) in the context of psychological research is, we suggest, likely to be a gross overestimate that will rarely be found in a large sample or in a replication. Smaller effect sizes are not merely worth taking seriously. They are also more believable.

Summary and Conclusion

This article began by describing problems with the traditional evaluation of effect sizes, including common ways in which they are misinterpreted – the most common mistake being to describe them in ways that either convey no useful information or are actively misleading. Next, the article outlined several ways (building on proposals by prior writers) to imbue effect size numbers with meaning. We concluded by offering some recommendations for the most useful ways to evaluate effect size and even, daringly, promulgated a new set of standards. Our hope is that this article might play a small role in helping to move effect sizes from numbers that are reported without interpretation, or interpreted superficially or incorrectly, to aspects of our research reports that will inform the application and theoretical development of psychological research.

[15] Notice that these benchmarks (except for .05) are the same as suggested by Gignac & Szodorai (2016), but with more generous labeling.

References

Abelson, R.P. (1985). A variance explanation paradox: When a little is a lot. *Psychological Bulletin, 97,* 129-133.

Ahadi, S., & Diener, E. (1989). Multiple determinants and effect size. *Journal of Personality and Social Psychology, 56,* 398-406.

Allen, M. S., & Walter, E. E. (2018). Linking big five personality traits to sexuality and sexual health: A meta-analytic review. *Psychological Bulletin, 144*(10), 1081-1110.

Aronson, E., & Carlsmith, J. M. (1968). Experimentation in social psychology. In G. Lindzey & E. Aronson (Eds.), The handbook of social psychology (2nd ed., Vol. 2, pp. 1-79). Reading, MA: Addison-Wesley.

Bargh, J.A., Chen, M., & Burrows, L. (1996). Automaticity of social behavior: Direct effects of trait construct and stereotype activation in on action. *Journal of Personality and Social Psychology, 71,* 230-244.

Baumeister, R.F., Vohs, K.D., & Funder, D.C. (2007). Psychology as the science of self-reports and finger movements. Whatever happened to actual behavior? *Perspectives on Psychological Science, 2,* 396-403.

Benjamin, D. J., Berger, J., Johannesson, M., Nosek, B. A., Wagenmakers, E., Berk, R., … Johnson, V. (2017, July 22). Redefine statistical significance. https://doi.org/10.31234/osf.io/mky9j

Camerer, C.F., Dreber, A., Holzmeister, F., Ho, T.H., Huber, J., Jahannesson, M., Kirchler, M., Nave, G., Nosek, B.A., Pfeiffer, T., Altmejd, A., Buttrick, N., Chan, T., Chen, Y., Forsell, E., Gampa, A., Heikensten, E., Hummer, L., Imai, T., Isaksson, S., Manfredi, D., Rose, J., Wagenmakers, E-J., & Wu, H. (2018). Evaluating the replicability of social science experiments in *Nature* and *Science* between 2010 and 2015. *Nature Human Behavior.* Published on-line 27 August 2018. Retrieved 28 August 2018 from https://www.nature.com/articles/s41562-018-0399-z.

Carroll, A.E., (2018, August 28). Study causes splash, but here's why you should stay calm on alcohol's risks. *New York Times.* Retrieved on-line, September, 6, 2018 from https://www.nytimes.com/2018/08/28/upshot/alcohol-health-risks-study-worry.html

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences* (rev. Ed.). New York: Academic Press.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd Ed.). Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Cohen, P., Cohen, J., Aiken, L.S., & West, S.G. (1999). The problem of units and the circumstances for POMP. *Multivariate Behavioral Research, 34,* 315-346.

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis.* New York: Taylor & Francis.

Darley, J.M., & Batson, C.D. (1967). "From Jerusalem to Jericho": A study of situational and dispositional variables in helping behavior. *Journal of Personality and Social Psychology, 27,* 100-108.

Darley, J.M., & Latané, B. (1968). Bystander intervention in emergencies: Diffusion of responsibility. *Journal of Personality and Social Psychology, 28,* 377-383.

Darlington, R.B. (1990). *Regression and linear models.* New York: McGraw-Hill.

De Boeck, P., & Jeon, M. (2018). Perceived crisis and reforms: Issues, explanations and remedies. *Psychological Bulletin, 144,* 757-777.

Doyen, S., Klein, O., Pichon, C-L., & Cleeremans, A. (2012). Behavioral Priming: It's All in the Mind, but Whose Mind? PLoS ONE 7(1): e29081. https://doi.org/10.1371/journal.pone.0029081\

Dweck, C. (2018, June 26). Growth mindset interventions yield impressive results. *The Conversation.* Download September 19, 2018 from https://theconversation.com/growth-mindset-interventions-yield-impressive-results-97423

Epstein, S. (1979). The stability of behavior: I. On predicting most of the people much of the time. *Journal of Personality and Social Psychology, 37*(7), 1097-1126. doi:http://dx.doi.org/10.1037/0022-3514.37.7.1097

Festinger, L., & Carlsmith, J.M. (1959). Cognitive consequences of forced compliance. *Journal of Abnormal and Social Psychology, 58,* 203-210.

Finkel, E.J., Eastwick, P.W., & Reis, H.T. (2017). Replicability and other features of a high-quality science: Toward a balanced and empirical approach. *Journal of Personality and Social Psychology, 113,* 244-253. doi:10.1037/pspi0000075

Fraley, R. C., & Marks, M. J. (2007). The null hypothesis significance testing debate and its implications for personality research. In R. W. Robins, R. C. Fraley, & R. F. Krueger (Eds.), *Handbook of research methods in personality psychology (pp. 149-169)*. New York: Guilford.

Friedman, H. S., Tucker, J. S., Tomlinson-Keasey, C., Schwartz, J. E., Wingard, D. L., & Criqui, M. H. (1993). Does childhood personality predict longevity? *Journal of Personality and Social Psychology, 65*, 176–185.

Funder, D.C. (2013). Does effect size matter? (Blog post). Retrieved on-line August 27, 2018 from https://funderstorms.wordpress.com/2013/02/01/does-effect-size-matter/.

Funder, D.C., & Ozer, D.J. (1983). Behavior as a function of the situation. *Journal of Personality and Social Psychology, 44,* 107-112.

Gelman, A. (2018, September 13). Discussion of effects of growth mindset: Let's not demand unrealistic effect sizes. *Statistical Modeling, Causal Inference, and Social Science* (blog). Retrieved September 19, 2018 from https://andrewgelman.com/2018/09/13/discussion-effects-growth-mindset-lets-not-demand-unrealistic-effect-sizes/

Gignac, G.E. & Szodorai, E.T. (2016). Effect size guidelines for individual differences researchers. *Personality and Individual Differences, 102,* 74-78.

Harding Center for Risk Literacy. (2018a) Fact boxes. Retrieved August 26, 2018 from https://www.harding-center.mpg.de/en/fact-boxes.

Harding Center for Risk Literacy (2018b). Antibiotics for acute bronchitis. Retrieved September 20, 2018 from https://www.harding-center.mpg.de/en/fact-boxes/use-of-antibiotics/acute-bronchitis.

International Situations Project (2018). Unpublished analyses. University of California, Riverside.

Jussim, L. (2017). Précis of *Social Perception and Social Reality: Why accuracy dominates bias and self-fulfilling prophecy. Behavioral and Brain Sciences, 40,* E1. doi:10.1017/S0140525X1500062X

Lakens, D., Scheel, A.M., & Isager, P.M. (2018). Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science, 1,* 259-269.

Meyer, G.J., Finn, S.E., Eyde, L.D., Kay, G.G., Moreland, K.L., Dies, R.R., Eisman, E.J., Kubiszyn, T.W., & Reed, G.M. (2001). Psychological testing and psychological assessment: A review of evidence and issues. *American Psychologist, 56,* 128-165.

Milgram, S. (1975). *Obedience to authority.* New York: Harper and Row.

Nisbett, R.E. (1980). The trait construct in lay and professional psychology. In L. Festinger (Ed.), *Retrospections on social psychology* (pp. 109-130). New York: Oxford University Press.

Ozer, D.J. (1985). Correlation and the coefficient of determination. *Psychological Bulletin, 97,* 307-315.

Ozer, D.J. (1993). Classical psychophysics and the assessment of agreement and accuracy in judgments of personality. *Journal of Personality, 61,* 739-767.

Ozer, D. J. (2007). Evaluating effect size in personality research. In R.W. Robins, R.C. Fraley, and R.F. Krueger (Eds.). *Handbook of Research Methods in Personality Psychology*. New York: Guilford.

Ozer, D.J., & Benet-Martínez, V. (2006). Personality and the prediction of consequential outcomes. *Annual Review of Psychology, 57,* 401-421.

Prentice, D.A., & Miller, D.T. (1992). When small effects are impressive. *Psychological Bulletin, 112,* 160-164.

Roberts, B. W., Kuncel, N. R., Shiner, R., Caspi, A., & Goldberg, L. R. (2007). The power of personality: The comparative validity of personality traits, socioeconomic status, and cognitive ability for predicting important life outcomes. *Perspectives on Psychological Science, 2*(4), 313–345.

Rosenthal, R. (1990). How are we doing in soft psychology? *American Psychologist, 45,* 775–777.

Rosenthal, R. (1996). Experimenter effects in behavioral research. In Rosenthal, R., and R.L. Rosnow, *Artifacts in Behavioral Research* (pp. 289-666). Oxford and New York: Oxford University Press.

Rosenthal, R., & Rubin, D.B. (1982). A simple, general purpose display of magnitude of experimental effect. *Journal of Educational Psychology, 74,* 166-169.

Schönbrodt, F.D., & Perugini, M. (2013). At what sample size do correlations stabilize? *Journal of Research in Personality, 47,* 609-612.

Sisk, V.F., Burgoyne, A.P., Sun, J., Butler, J.L., & Macnamara, B.N. (2018). Two what extent and under which circumstances are growth mind-sets important to academic achievement? *Psychological Science,* 29(4), 549-571.

Vohs, K. (Chair) (2018, March). *A pre-registered depletion replication project: The paradigmatic replication approach.* Symposium at the Society for Personality and Social Psychology, Atlanta.

Weston, S.J., Gladstone, J.J., Graham, E.K., Mroczek, D.K., & Condon, D.M. (2018). Who are the scrooges? Personality predictors of holiday spending. *Social Psychological and Personality Science.* Published on-line in advance of print. Retrieved October 9, 2018 from http://journals.sagepub.com/doi/full/10.1177/1948550618792883.