aps ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Perspectives on Psychological Science
1–9

DOI: 10.1177/1745691620970557
www.psychologicalscience.org/PPS
SAGE

When the Numbers Do Not Add Up: The Practical Limits of Stochastologicals for Soft Psychology

Nick J. Broers
Department of Methodology and Statistics, Faculty of Psychology and Neuroscience, Maastricht University

Abstract

One particular weakness of psychology that was left implicit by Meehl is the fact that psychological theories tend to be verbal theories, permitting at best ordinal predictions. Such predictions do not enable the high-risk tests that would strengthen our belief in the verisimilitude of theories but instead lead to the practice of null-hypothesis significance testing, a practice Meehl believed to be a major reason for the slow theoretical progress of soft psychology. The rising popularity of meta-analysis has led some to argue that we should move away from significance testing and focus on the size and stability of effects instead. Proponents of this reform assume that a greater emphasis on quantity can help psychology to develop a cumulative body of knowledge. The crucial question in this endeavor is whether the resulting numbers really have theoretical meaning. Psychological science lacks an undisputed, preexisting domain of observations analogous to the observations in the space-time continuum in physics. It is argued that, for this reason, effect sizes do not really exist independently of the adopted research design that led to their manifestation. Consequently, they can have no bearing on the verisimilitude of a theory.

Keywords

effect size, verbal versus formal theories, ordinal versus quantitative predictions, meta-analysis

If there is one key excerpt from Meehl's (1978) seminal article that could serve as a careful reflection on the replication crisis in psychology, it is surely this one:

> Meehl's Mental Measure correlates .50 with SES in Duluth junior high school students, as predicted from Fisbee's theory of sociability. When Jones tries to replicate the finding on Chicano seniors in Tucson, he gets $r = .34$. Who can say anything theoretically cogent about this difference? Does any sane psychologist believe that one can do much more than shrug? (p. 814)

What Meehl alerted us to with this example is that psychologists are not dealing with nomological relationships of the type that we encounter in the natural sciences but rather with probabilistic sort of equivalents that may vary in magnitude from one context to another, without any psychologist being able to explain the reason for the difference. Meehl introduced the term *stochastological* to signify the probabilistic lawlike relationships that litter the theoretical landscape of soft psychology. The term never caught on. We use the older and more familiar "effect size" instead to denote the same lawful relationship for meta-analytic purposes often expressed in the form of a correlation coefficient. Strikingly, Meehl (1978) observed that "thoughtful theorists realize how little *quantitatively* we can say with sufficient confidence to warrant counting an unexpected shift in a stochastic quantity as a strong 'discorroborator'" (p. 814). Given this profoundly consequential observation, it is tempting to speculate how Meehl would have reacted to the much discussed failed attempts at replication more than 35 years later (Marsman et al., 2017; Open Science Collaboration, 2015; Pashler & Wagenmakers, 2012). He probably would have been neither impressed nor shocked, particularly because he

Corresponding Author:
Nick J. Broers, Department of Methodology and Statistics, Faculty of Psychology and Neuroscience, Maastricht University
E-mail: nick.broers@maastrichtuniversity.nl

did not think highly of the practice of null-hypothesis significance testing (NHST).

Like Tukey (1991) and J. Cohen (1994), Meehl felt that even in the case of randomized experiments, the null hypothesis was bound to be false anyway and that for this reason a nonsignificant result ultimately signified a lack of power to establish the sign of the existing nonzero effect. The prediction that was falsified (a zero effect) was uninteresting to begin with, and therefore the ritual of NHST constitutes bad science. But of course this is not how many researchers tend to perceive the significant outcome of the inferential test. Although few would be so naive as to contend that the rejection of the null hypothesis implies the truth of the alternative, the practical rationale for the NHST procedure is to confirm the theoretically based prediction. Taking a Popperian perspective, this traditional procedure could lead only to sterile science according to Meehl. We cannot confirm the truth of a scientific theory, but we can increase support for its "verisimilitude" or truthlikeness by subjecting the theory to a collection of ever increasing high-risk tests.

Buried in his critique of theoretical progress in soft psychology is one inherent weakness in theoretical psychology, largely left implicit by Meehl: Soft psychology makes use of verbal rather than of formal theories of behavior. Verbal theories do not enable theorists to make quantitative predictions of the sort that would yield the risky test (i.e., tests with a low a priori success probability from a Popperian point of view) that strengthens belief in the verisimilitude of a theory. Meehl's own example of rain in April provides a case in point. The uninteresting hypotheses—that it will rain in April and that it will rain more so than in May—can be derived from a verbal theory, but all of the more interesting hypotheses (starting with the prediction that it will rain exactly 7 days in April) cannot.

With the advent of meta-analyses in the 1980s, psychologists tried to circumvent this limitation of verbal theories by capturing quantities in the wild rather than by deriving them from tested theory. The observed effect size yielded by experiment or uncovered in observational research is believed to have theoretical meaning. For this reason, the statistical practice of meta-analysis is thought to help us uncover the true size of the theoretical effect or to pinpoint the way to interesting moderators of the observed effect. But this practice amounts to the theoretical derivation of the hypothesis that it will rain in April (a qualitative prediction), supplemented by the statistical revelation that, on average, it will rain 7 days in April (a quantitative finding). The latter observation, although potentially interesting, is not rooted in theory. As I argue, that poses a problem for soft psychology.

Quantity in Psychology

Ever since Glass (1976) first introduced the term "meta-analysis," this technique for combining the results of multiple studies has been applied so ubiquitously in all psychological disciplines that it has since been followed by meta-meta-analysis studies. In a summary of 100 years of social-psychological research covering more than 25,000 research studies, Richard et al. (2003) presented a quantitative overview of the magnitude and variability of social-psychological effects. They established that roughly 30% of the reported effects were small (around $r = 0.10$), approximately half of the reported effects were around 0.20, and less than 25% were found to be larger than 0.30.

Hedges (1987) claimed that the growing emphasis on quantification through the use of effect-size estimates decreases the hierarchical difference between hard physical science and soft psychological science.[1] Hedges conceded that there is a clear difference in theoretical cumulativeness between psychological and physical science. Over the centuries, physics has succeeded into grouping simple laws within ever more general laws that build on each other to yield ever more refined insights into the details of physical reality. In stark contrast, psychological science has remained stuck in comparatively simple explanatory principles that so far could not be expanded into a hierarchical theoretical structure. However, a much more favorable comparison results when examining the empirical cumulativeness of both physical and psychological science. With the term *empirical cumulativeness*, Hedges referred to the level of agreement among replicated studies, as reflected by the outcomes of meta-analyses.

In his research, Hedges (1987) sampled 13 study reviews from particle physics (comparable to a meta-analysis in social science) with the aim of estimating either mass or lifetime of a wide collection of subatomic particles. The null hypothesis that the estimates showed no systematic variability across studies had to be rejected for six of the 13 studies (i.e., when all studies were included in the review). (A subset of studies was actually not included on the basis of several common exclusion criteria within particle physics. For the remaining studies, only two led to a rejection of the null hypothesis of no systematic variability across studies). Hedges next sampled 13 meta-analyses that either dealt with sex differences in cognitive abilities or with the effects of different educational programs on academic achievement. All individual studies yielded effects that were quantified with the standardized Cohen's *d* measure. The null hypothesis that these *d* estimates showed no systematic variability across studies was rejected in six of the 13 meta-analyses (or in none of these 13 studies if some

studies were left out on the basis of a commonly accepted exclusion criterion). Hedges concluded that the empirical replicability of effects is actually quite similar across the two scientific disciplines. In terms of empirical cumulativeness at least, psychology is not demonstrably "softer" than physics.

The growing popularity of meta-analyses seems to have put psychology on a course to become a truly quantitative science, building cumulative empirical knowledge (Hedges, 1987). Does this mean that psychology has found a wormhole that takes it from soft verbal theorizing straight to hard quantitative knowledge? And does this imaginative shortcut to seemingly stable and hard knowledge provide support for the verisimilitude of the underlying theories? Some methodological reformists seem to think so. Proponents of the so-called New Statistics have expressed dissatisfaction with the fact that traditional significance testing tends to produce dichotomous (effect vs. no effect) outcomes; they believe these outcomes do not really provide a deepening of understanding or a gradual accumulation of research insights. Instead, they advocate a methodological shift to the estimation of the size and stability of psychological effects (Cumming, 2014). The growing influence of this movement is apparent from the change in editorial policy of *Psychological Science* (Eich, 2014) and especially from the truly revolutionary decision of the editorial board of *Basic and Applied Social Psychology* to dispense with NHST altogether (Trafimow & Marks, 2015).

The pivotal role of effect size in this endeavor deserves a more critical examination of this concept than is common within the current scientific climate. The key question must be how a scientific discipline that rests on purely verbal theories succeeds in inserting quantity into its thinking. The important follow-up question must be what meaning, if any, do such quantities have at the theoretical level of psychology.

To most psychologists, the question of just how quantity is being inserted in their specific discipline seems deceptively simple to answer. Quantity emerges as soon as participants in a psychological experiment are scored on the outcome variable. In many if not most studies, the outcome variable is measured with a rating scale or by determining the sum or the average on a collection of related Likert scales (P. Cohen et al., 1999). The information thus obtained is assumed to be of a quantitative nature—assumed because with few exceptions (e.g., Borsboom & Mellenbergh, 2004) psychologists do not actually test whether the data satisfy the necessary conditions for establishing quantitative measurement (Michell, 2008). Instead, most researchers rely on classical test theory to determine unidimensionality and internal consistency. Although this forms the traditional procedure for determining scalability, it does not provide a formal test of a quantitative data structure. Such a test could be achieved by the application of additive conjoint measurement theory (Luce & Tukey, 1964), but this is rarely done in practice (Cliff, 1992). Instead of carrying out such a formal test, psychologists use Stevens's classification of measurement scales (Stevens, 1946) and choose the scale type that they believe captures the properties of their measurement instrument. On the basis of the untested assumption that we are dealing with quantitative data as opposed to merely ordinal data, psychologists then usually proceed to make use of parametric statistical models to test hypotheses on the equality of population means. If we are willing to consider it as plausible that the assumption of quantitative data is actually met, as most psychologists do, the habitual use of inferential statistics by itself seems to qualify psychology as a quantitative scientific discipline. According to Michell (2008), this conscious neglect of the scientific task of demonstrating that the data do actually exhibit the properties of a quantitative structure turns psychology into pathological science. However, as I argue, the shift from NHST to meta-analysis and effect-size estimation will turn the diagnosed pathology into a more acute problem.

Verbal Theories and Ordinal Predictions

Commenting on Thomas Kuhn's famous concept of paradigm shifts in scientific revolutions (Kuhn, 1962), Nobel laureate in physics Steven Weinberg stressed that a distinction should be made between the hard parts and the soft parts of a physical theory:

> There is a "hard" part of modern physical theories . . . that usually consists of the equations themselves, together with some understandings about what the symbols mean operationally and about the sorts of phenomena to which they apply. Then there is a "soft" part; it is the vision of reality that we use to explain to ourselves why the equations work. The soft part does change; we no longer believe in Maxwell's ether, and we know that there is more to nature than Newton's particles and forces. (Weinberg, as quoted in Van Fraassen, 2001, p. 163)

The theoretical rationales for understanding relationships between phenomena in the physical domain change, but the equations that describe these relationships remain the same: The acceleration of a body in free fall remains 9.8 m/s^2 irrespective of whether we use Newton's or Einstein's perspective of the universe for explaining it.

Although from a philosophical point of view it can be argued that all observation is theory-laden (e.g.,

Bridgman, 1927), from a practical perspective this is more evidently true for psychology than for the natural sciences. In psychology the distinction between hard and soft parts of theories is decidedly less clear because of the gap between the theoretical concepts and their empirical operationalizations. In physics the theoretical constructs are directly rooted in the empirical relationships. They are represented by parameters in formal models that predict and describe structures in the empirical observations (Margeneau, 1950). Theories in physics pertain to structural regularities that can be plotted in the space-time continuum. As the above quote from Weinberg underlines, these regularities have an existence independent of the theory that describes them. This means that these regularities already existed before theoretical consideration led to their prediction. The domain of observations to which the theory pertains are certain empirical regularities in the space-time continuum. In contrast, the theories in psychology do not pertain to an independent and objective domain of observations. With the notable exception of laws in psychophysics (e.g., Fechner's law) or laws that apply to motor behavior (e.g., Fitts's law), there are only a few empirical regularities of a quantitative nature waiting for an appropriate psychological theory to uncover them.

Instead, the domain of observations to which the theory pertains is a diffuse collection of qualitative psychological experiences. Festinger, for instance, was first triggered to theorize about cognitive dissonance on the basis of his experiences with the doomsday cult that prophesied the end of the world and whose members intensified their beliefs when this did not happen (Festinger et al., 1956). Such an experience can be psychologically explained with dissonance theory, which subsequently is believed to shine a light on different but related phenomena, such as why people show more appreciation of a course if they had to pay a lot of money or exert a lot of effort to take it. It turns out that a potent theory such as cognitive dissonance has something to tell about a large diversity of psychological experiences but not about any quantitative relationship in a preexisting domain of observations. The diffuse collection of psychological experiences that gives rise to psychological theories is inherently qualitative, not quantitative.[2] Quantity comes into play only as soon as the theorist has devised a research design that should enable a qualitative verdict on the theory: An ordinal relationship has been shown to occur (e.g., the group that received an experimental treatment scores significantly higher on a rating scale than a control group that did not receive treatment) or not (the difference is not statistically significant). To enable significance testing, some scale needs to be devised that will produce scale values that can be used for computing means and standard deviations. These statistics help to establish an observed effect size, but that effect size, it is crucial to realize, did not have an independent existence in nature before the theorist constructed this research design as a means for eliciting a predicted (ordinal) effect.

In addition, and perhaps more importantly, the theory never *predicted* the occurrence of any quantitative relationship. Most theories in psychology are verbal theories rather than formal or computational ones.[3] They provide a narrative exposition of how an hypothesized psychological mechanism is thought to operate. This gives the theorist an intuitive understanding of what is going on and what may be expected in an empirical study, but the reliance on language is inevitably accompanied by vagueness and imprecision, both in regard to the meaning of the theoretical concepts and the functional relationships between these concepts (Dennis & Kintsch, 2007; Hintzman, 1991).

As a consequence, predictions made by a verbal theory can be only ordinal at most: The group receiving Treatment A will produce a higher mean on the outcome scale than the group receiving Treatment B. There is no theoretical basis for a prediction such as "the mean of A will be 5 points higher than that of B."[4] This means that the observed effect size does not actually provide any information about the theory. If A was predicted to be greater than B, then actually observing this relationship provides empirical support for the theory. Many replications of the same result likewise provide information about the theory because these replications strengthen its credibility. However, the quantitative result (i.e., the magnitude of the observed effect) in any particular study forms at best an explanandum: It is an unexpected result that we might wish to explain in a more formal development of the theory. But this is unlikely to be the case if the outcome scale is not inherently meaningful. We must therefore conclude that the observed effect size in any individual study does not give us any information about the theory that the study aimed to test. It is only when further theoretical development enables us to explain the *quantitative* result (as opposed to merely explaining the ordinal result) that the effect size actually obtains theoretical meaning. Only then can an empirically observed quantity lend support for the verisimilitude of a theory. If the quantity in any given research design is itself considered as interesting because the outcome scale is considered as meaningful and should be retained in future studies on the same phenomenon, then the observed effect size is taken as an explanandum, and future effort will be directed toward developing a formal theory to explain this result.

Qualitative Understanding of Quantitative Findings

Meehl (1978) mentioned 20 difficulties that are inherent in psychological science. At least three of these are directly related to the current analysis. A key problem that Meehl mentioned is the fact that theoretical concepts in psychology tend to be open concepts. Precisely because they originate in the experiential world of everyday life, psychological concepts are inherently rich in meaning and fuzzy in nature. Any concrete operationalization enables the researcher to focus on the concept in a carefully defined and quantifiable way but inevitably at the cost of simplification. As Meehl observed, the provisional list of operational indicators of a concept is indefinitely extensible. The inevitable theoretical surplus meaning of psychological concepts may well be one of the major reasons why psychological theories have remained verbal theories. The narrative structure of such theories makes it easier to incorporate the open meaning of the concepts than mathematical formalism would.

Testing hypotheses derived from verbal theories introduces a difficulty that Meehl called the "response-class problem," referring to the difficulty of determining which particular attributes should be chosen for establishing the outcome of the experimental manipulation, as well as determining how these should be "sliced up" to provide meaningful outcome measures. When testing an explanatory psychological mechanism such as cognitive dissonance, the halo effect, or conformity, there exists an undetermined wealth of possible ways for establishing a predicted effect. Likewise, there is a similar unspecified multitude of different ways in which researchers can elicit the hypothesized effect, a problem that Meehl referred to as the "situation-taxonomy problem." The number of different settings and stimuli that researchers can use for testing hypotheses is restricted only by their imagination. There is obviously no objective basis for expecting an observed effect size found on a particular outcome scale in one setting with one set of stimuli to be exactly or even roughly equal to the size of the same effect expressed on a different outcome scale in another setting with different stimuli. Other than a semantic association there is no basis for such an expectation because although theoretically we may be dealing with an expression of the same explanatory mechanism in both studies, our verbal theory does not allow any quantitative prediction for either study, so an explanation for any difference between the two cannot be given either.

Mook (1983) argued that this freedom in choosing settings for and ways of demonstrating theoretical effects does not invalidate or diminish the scientific findings but conversely may even strengthen them. In his "defense of external invalidity," Mook claimed that the preoccupation with settings and manipulations arises only when one wants to generalize the specific finding in the laboratory to the mundane world. However, in theoretical research we are often interested in studying the workings of a proposed explanatory psychological mechanism. We want to gain greater understanding of how this mechanism operates. It is this understanding that we wish to generalize, not the actual findings themselves. For Mook, the artificiality of laboratory settings (and the accompanying low external validity) is not a problem because we are not primarily interested in predicting how people will be acting in the real world but instead in gaining an understanding of the way in which one of the determinants of that behavior operates behind the scenes.

Mook concluded that "Ultimately, what makes research findings of interest is that they help us understand everyday life. That understanding, however, comes from theory or the analysis of mechanism; it is not a matter of 'generalizing' the findings themselves" (p. 386). The verbal theories of psychology seek to further our understanding through the proposition of theoretical mechanisms. Clever experiments elicit the predicted ordinal effects through the use of arbitrary settings, stimuli and outcome scales, yielding mean differences and correlations that have no intrinsic meaning but that do allow us to draw conclusions on the working of the proposed mechanisms. It is this *qualitative* understanding that can be transferred to nonartificial real-world situations, in which meaningful outcomes can be registered and reported. In these practical contexts, the observed *quantitative* effect size has or can have real meaning. But that effect size is not explained by the theory. The only thing that is explained by theory is the effect itself.

Cataloguing Swans

Although Meehl's listing of 20 difficulties peculiar to psychology (Meehl, 1978) makes plain that psychologists cannot hope to build a cumulative theoretical structure of the sort that helped to produce rocket science, it is clear that he believes that true progress will nonetheless require psychological theories to be tested in a way that will enable us to establish support for their verisimilitude—an aspiration that is not likely to be met by the traditional practice of NHST. But will the increasing emphasis on effect size and meta-analysis help us to achieve this goal?

Whereas NHST leads to the falsification of an hypothesis that is not very challenging from the outset, estimation and subsequent meta-analysis of effect sizes form

an endeavor that seems focused on cataloguing swans rather than on putting theories to a high-risk test. Bear in mind that the verbal psychological theory that predicted the effect never predicted the observed *effect size* that was registered across these studies. Finding that a collection of studies on a single psychological phenomenon produced effect-size estimates that can be described by a homogeneous-effects model therefore seems like concluding that only white swans have been observed in the theoretical landscape. That is an observation, not a theoretically based prediction. At a later stage further studies on the same phenomenon eventually produce larger effects, and a new meta-analysis now rejects a homogeneous-effects model in favor of a random-effects model: Instead of only white swans, some black ones have unexpectedly turned up. The introduction of methodological or substantive moderators may help us to restore some order again: Within strata all swans are white again—until eventually some black ones pop up once again, requiring the search for further moderators.

I think it is evident that a shift from NHST to the estimation of effect sizes is not going to help us much further this way. If we cannot predict a quantity, whatever quantity is turning up in a single study or in a collection of direct or conceptual replications of the study will never allow us to gauge the verisimilitude of the underlying theory. Let us observe this problem from the opposite angle. If a meta-analysis uncovers the common size of the effect (r) to be .20, exactly what sort of prediction does that allow us to make on the domain of psychological observations? Because this domain is essentially qualitative and cannot be meaningfully defined or demarcated, the question cannot be answered. The only prediction we can make is the ordinal prediction that the likely effect to be produced will be relatively small, a prediction that is based on theory (there will be an effect) as well as on statistics (this was found to be small across different studies). If you have seen only white swans, the expected outcome will be for the next one to turn up white as well, without understanding why this should be the case. Again, the key problem is the lack of a naturally preexisting domain of observations to which psychological theories pertain. In psychology there are no black holes or gravity waves just waiting to be discovered by the appropriate theory.

Glöckner (2016) presents an interesting example that illustrates this nicely. He discussed a study by Danziger et al. (2011) that found that judges handling a sequence of parole cases tend to pass more lenient judgments at the beginning of a session but typically more severe judgments toward the end. The authors speculated that the finding might be (partially) explained as the effect of mental depletion. However, the effect that they found was conspicuously large (converted into a Cohen's d = 1.96), which seemed at odds with meta-analytic findings that suggested the mental-depletion effect to be rather small (with a publication-bias corrected interval estimate for d between −0.10 and 0.25; see Carter & McCullough, 2013). However, these d values were all catalogued in laboratory studies that typically make use of stimuli and outcome measures with little relevance in the real world. Martijn et al. (2007), for example, used difficult puzzles to mentally deplete their subjects, and the effect was then registered by having them squeeze a handgrip for as long as they could. As Glöckner observed, the depletion effect in the courtroom was possibly much larger than typically found because in this case we were dealing with a real-world setting as opposed to the laboratory studies that yielded estimates of a small effect.[5] This amounts to proposing a methodological moderator for the unexpected observed heterogeneity in effects. But if we had no theory to explain why the mental-depletion effect would be around d = 0.08 or even simply why it would be small, how can we ever use the outcome of mental-depletion studies for predicting the size of the effect in a domain that is so very different from what was observed in the laboratory? As Meehl observed almost 40 years ago with regard to context-dependent stochastologicals, if the size of the effect turns out to be different from what was observed before, can we do anything but shrug?

Conclusion

This study was motivated by a desire to better understand the status of effect size in psychological research and was guided by two leading questions. First, how does academic psychology, which rests largely on verbal theories, succeed in inserting quantity into its thinking, and second, what meaning do such quantities have at the theoretical level of psychology?

Put crudely, the answer to the first question is actually by an article of faith. Researchers in psychology ubiquitously make use of rating scales and other measurement instruments that are treated as interval scales without attempting to test whether the data really do exhibit quantitative properties. But although the status of the assumed quantities is habitually debatable (e.g., "Are these data truly measured on an interval scale or merely ordinal?" "Can we trust the means and standard deviations to have true quantitative meaning here?" "Are these truly linear correlations or perhaps more like rank-order correlations?"), that need not be of overriding concern as long as the focus is primarily on the sort of hypothesized ordinal relationships that can be derived from a verbal theory. The use of parametric statistics requires quantitative data (and some assumptions on the structure of the data), but if the researcher and intended audience

are willing to take that for granted, then finding a mean difference to be significant leads us to conclude that support has been found for the interesting theory-based prediction that A is greater than B.

The second question on the meaning of the quantities comes into play only when we shift our attention from the ordinal finding that was predicted by the theory to the quantitative finding that was not. As soon as we attempt to quantify the size of the effect the stakes become higher. Take for example the famous dissonance study by Festinger and Carlsmith (1959) on the appreciation of a boring task after being given either a small or large reward to lie about its interestingness. The outcome was measured on an 11-point rating scale (−5 to +5), the mean difference between the two groups was 1.40, and the common standard deviation was estimated as 2. The metric of the outcome scale was probably chosen for its intuitive appeal but otherwise strictly arbitrary. What does the mean difference of 1.40 express? It is an unbiased estimate of the true mean difference in the population of similar participants being exposed to either the $1 or $20 reward manipulation and being measured on this 11-point outcome scale. Obviously nobody would consider that an interesting revelation. Now suppose we have a rival theory that proposes an entirely different psychological mechanism for explaining why participants will on average show more enjoyment of the boring task in the $1 condition. The rival theory makes the same ordinal prediction, so the mean difference found in the original study lends equal support to both theories. The observed effect size was not predicted by either theory and so cannot be of use to favor one theory over the other. In other words, quantitative effect sizes in psychology must be considered as theoretical orphans.

At around the same time as the Festinger and Carlsmith study, another important study on the effect of cognitive dissonance was conducted by Aronson and Mills (1959),[6] who showed that the appreciation of a boring discussion group was affected by the severity of the initiation procedure that preceded it. In their study, appreciation was measured on a 16-point rating scale (from 0 to 15), and the difference between the group that had a mild initiation and the group that had a severe initiation was 1.60. Again, the arbitrariness of the scaling units turns this into a meaningless number, which is why it would traditionally be converted into Cohen's *d*. It turns out that this Cohen's *d* equals 0.73, which is quite close to the value of 0.70 that was found for the Festinger and Carlsmith study.

It is tempting to believe that the close correspondence of the results provides us with real quantitative information on the effect of dissonance reduction. But it is important to realize here that the verbal theory that inspired the experiments not only had nothing quantitatively to say about the workings of cognitive dissonance but also was not developed on the basis of any quantitative observation in the first place. Thinking about cognitive dissonance was developed on the basis of a large collection of essentially qualitative experiences covering a wide range of phenomena and situational contexts. There was nothing quantitative to explain, and the arbitrariness of the outcome scales reflects the indifference of the researchers toward whatever quantitative outcome the study might yield. The only purpose of quantification was to enable an NHST ritual to underwrite the ordinal theoretical prediction. The conclusion must then be that observed effect sizes have no meaning outside the research design in which they were established. Consequently, they can have no bearing on the verisimilitude of theories. Those who, like proponents of the New Statistics movement, nonetheless want to collect all of these quantities in the wild and use them as a source of information on what is happening at the theoretical landscape should at least attempt to demonstrate that the effects found in individual studies really reflect quantitative information. Failure to do so makes the accusation of pathological science (Michell, 2008) more serious than in the case of NHST, in which the primary objective is to provide empirical support for essentially ordinal predictions.

Because effect size does have meaning within the confines of a chosen research design, the reporting of an observed effect size obviously provides important supplementary information to the outcome of the significance test. Measuring outcomes on a scale with meaningful units will allow us to consider the psychological importance of the observed result. Presenting a result will not be informative until we can gauge its psychological relevance, especially in an applied context. An example is provided by Axsom and Cooper (1985), who used a bogus treatment for losing weight to demonstrate that dissonance reduction could be responsible for significant weight loss in a group of overweight women. The outcome of their study showed that high effort justification did indeed result in a durable mean weight loss of around 8 lb but that low effort justification resulted in a negligible weight loss of somewhere between 0 and 1 lb.

In a theoretical context it may sometimes be useful to determine the smallest effect of interest. Although the examples presented in Lakens et al. (2018) show that such smallest effects are usually derived on the basis of subjective considerations, there are instances in which computational theories permit the quantitative prediction of such an effect. In addition, there are rare

instances in which even a verbal theory provides clues that suggest a smallest effect of interest (see Burriss et al., 2015).

However, as I have argued in this article, the stochastological use of effect size is of little use because the quantities have no real theoretical meaning. Cumulative development of psychological knowledge will not manifest itself as an increase of quantitative understanding because there are essentially no quantities to understand. The fact that a meta-analysis shows us that the average effect size of a psychological effect was small but that some individual studies produced large effects cannot be interpreted as evidence that the true size of the effect is probably small because there is no true magnitude to be discovered in the first place. Such a result of a meta-analysis tells us at best only that many of the reported studies failed to produce psychologically relevant demonstrations of the effect, but some studies did. The true accumulation of theoretical knowledge lies in a gradual increase of the practical relevance of the theoretical psychological principles. As the number of successful conceptual replications of a study multiplies, the breadth of applicability of these psychological principles will gradually increase, making the underlying theory both more potent and convincing. It will not take us to the moon, but it creates new and exciting perspectives on human behavior that satisfy our thirst for understanding.

Transparency

Action Editor: Travis Proulx and Richard Morey
Advisory Editor: Richard Lucas
Editor: Laura A. King
Declaration of Conflicting Interests
The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

ORCID iD

Nick J. Broers https://orcid.org/0000-0003-0672-9508

Notes

1. All references to "psychology" or "psychological science" throughout the remainder of the article pertain exclusively to the subdomains of psychology that traditionally work with verbal theories. It is the part of psychological science that is generally referred to as "soft" psychology.
2. This assertion should not be misunderstood as an ontological claim that psychological phenomena cannot be regressed to some form of objective, "hard" ultimate reality embedded in the brain or more generally in the biological system. However, the corresponding domain of observations is more likely to be the focus of cognitive neuroscience than that of psychology.
3. Formal theories pertain to mathematical models of behavior that allow the derivation of quantitative predictions. Mathematical formalizations of psychological theories have many advantages apart from the ability to make quantitative predictions. Most notably, they force theorists to make underlying assumptions explicit and impose consistency on reasoning. In addition, mathematical formalizations force researchers to be precise in defining theoretical concepts as well as the rules that govern their interaction (Hintzman, 1991).
4. Velicer et al. (2008) claimed to have succeeded in making quantitative predictions of effect size that were derived from a verbal theory. However, the predictions were in fact ordinal ("for this we expect a small effect size," "for that we expect a large effect size," etc.); they were turned into quantitative predictions with reference to conventional ω^2 values provided by J. Cohen (1988), which were then "recalibrated" on the basis of observed effect sizes in two previous studies. The predictions were based on an intuitive line of reasoning, and the way in which the traditional effect-size values were recalibrated was not made explicit.
5. Glöckner (2016) subsequently proceeded to test an alternative, more likely explanation for the surprising effect.
6. Both of these studies were discussed by Stukas and Cumming (2014), who advocated for a more quantitative approach to social psychology.

References

Aronson, E., & Mills, J. (1959). The effect of severity of initiation on liking for a group. *Journal of Abnormal and Social Psychology*, *59*, 177–181. https://doi.org/10.1037/h0047195

Axsom, D., & Cooper, J. (1985). Cognitive dissonance and psychotherapy: The role of effort justification in inducing weight loss. *Journal of Experimental Social Psychology*, *21*, 149–160. https://doi.org/10.1016/0022-1031(85)90012-5

Borsboom, D., & Mellenbergh, G. J. (2004). Why psychometrics is not pathological. A comment on Michell. *Theory & Psychology*, *14*, 105–120. https://doi.org/10.1177/0959354304040200

Bridgman, P. W. (1927). *The logic of modern physics*. Macmillan.

Burriss, R. P., Troscianko, J., Lovell, P. G., Fulford, A. J. C., Stevens, M., Quigley, R., Payne, J., Saxton, T., & Rowland, H. M. (2015). Changes in women's facial skin color over the ovulatory cycle are not detectable by the human visual system. *PLOS ONE*, *10*(7), Article e0130093. https://doi.org/10.1371/journal.pone.0130093

Carter, E. C., & McCullough, M. E. (2013). Is ego depletion too incredible? Evidence for the overestimation of the depletion effect. *Behavioral and Brain Sciences*, *36*, 683–684. https://doi.org/10.1017/S0140525X13000952

Cliff, N. (1992). Abstract measurement theory and the revolution that never happened. *Psychological Science*, *3*, 186–190. https://doi.org/10.1111/j.1467-9280.1992.tb00024

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Erlbaum.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, *49*, 997–1003. https://doi.org/10.1037/0003-066X.49.12.997

Cohen, P., Cohen, J., Aiken, L. S., & West, S. G. (1999). The problem of units and the circumstance for POMP.

Multivariate Behavioral Research, *34*, 315–346. https://doi.org/10.1207/S15327906MBR3403_2

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, *25*, 7–29. https://doi.org/10.1177/0956797613504966

Danziger, S., Levav, J., & Avnaim-Pesso, L. (2011). Extraneous factors in judicial decisions. *Proceedings of the National Academy of Sciences, USA*, *108*, 6889–6892. https://doi.org/10.1073/pnas.1018033108

Dennis, S., & Kintsch, W. (2007). Evaluating theories. In R. J. Sternberg, H. L. Roediger III, & D. F. Halpern (Eds.), *Critical thinking in psychology* (pp. 143–159). Cambridge University Press.

Eich, E. (2014). Business not as usual [Editorial]. *Psychological Science*, *25*, 3–6. https://doi.org/10.1177/0956797613512465

Festinger, L., & Carlsmith, J. M. (1959). Cognitive consequences of forced compliance. *Journal of Abnormal and Social Psychology*, *58*(2), 203–210. https://doi.org/10.1037/h0041593

Festinger, L., Riecken, H. W., & Schachter, S. (1956). *When prophecy fails*. University of Minnesota Press.

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher*, *5*, 3–8. https://doi.org/10.3102/0013189X005010003

Glöckner, A. (2016). The irrational hungry judge effect revisited: Simulations reveal that the magnitude of the effect is overestimated. *Judgment and Decision Making*, *11*, 601–610.

Hedges, L. V. (1987). How hard is hard science, how soft is soft science—The empirical cumulativeness of research. *American Psychologist*, *42*, 443–455. https://doi.org/10.1037/0003-066X.42.5.443

Hintzman, D. L. (1991). Why are formal models useful in psychology? In W. E. Hockley & S. Lewandowsky (Eds.), *Relating theory and data: Essays on human memory in honor of Bennet B. Murdock* (pp. 39–56). Erlbaum.

Kuhn, T. S. (1962). *The structure of scientific revolutions*. University of Chicago Press.

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. https://doi.org/10.1177/2515245918770963

Luce, R. D., & Tukey, J. W. (1964). Simultaneous conjoint measurement: A new scale type of fundamental measurement. *Journal of Mathematical Psychology*, *1*, 1–27. https://doi.org/10.1016/0022-2496(64)90015-X

Margeneau, H. (1950). *The nature of physical reality*. McGraw-Hill.

Marsman, M., Schönbrodt, F. D., Morey, R. D., Yao, Y., Gelman, A., & Wagenmakers, E.-J. (2017). A Bayesian bird's eye view of 'replications of important results in social psychology'. *Royal Society Open Science*, *4*, Article 160426. https://doi.org/10.1098/rsos.160426

Martijn, C., Alberts, H. J. E. M., Merckelbach, H., Havermans, R., Huijts, A., & De Vries, N. K. (2007). Overcoming ego-depletion: The influence of exemplar priming on self-control performance. *European Journal of Social Psychology*, *37*, 231–238. https://doi.org/10.1002/ejsp.350

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

Michell, J. (2008). Is psychometrics pathological science? *Measurement*, *6*, 7–24. https://doi.org/10.1080/15366360802035489

Mook, D. G. (1983). In defense of external invalidity. *American Psychologist*, *38*, 379–387. https://doi.org/10.1037/0003-066X.38.4.379

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*, Article aac4716. https://doi.org/10.1126/science.aac4716

Pashler, H., & Wagenmakers, E.-J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science*, *7*, 528–530. https://doi.org/10.1177/1745691612465253

Richard, F. D., Bond, C. F., Jr., & Stokes-Zoota, J. J. (2003). One hundred years of social psychology quantitatively described. *Review of General Psychology*, *7*, 331–363. https://doi.org/10.1037/1089-2680.7.4.331

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, *103*, 667–680.

Stukas, A. A., & Cumming, G. (2014). Interpreting effect sizes: Toward a quantitative cumulative social psychology. *European Journal of Social Psychology*, *44*, 711–722. https://doi.org/10.1002/ejsp.2019

Trafimow, D., & Marks, M. (2015). Editorial. *Basic and Applied Social Psychology*, *37*, 1–2. https://doi.org/10.1080/01973533.2014.865505

Tukey, J. W. (1991). The philosophy of multiple comparisons. *Statistical Science*, *6*, 100–116.

Van Fraassen, B. C. (2001). Constructive empiricism now. *Philosophical Studies*, *106*, 151–170. https://doi.org/10.1023/A:1013126824473

Velicer, W. F., Cumming, G., Fava, J. L., Rossi, J. S., Prochaska, J. O., & Johnson, J. (2008). Theory testing using quantitative predictions of effect size. *Applied Psychology*, *57*, 589–608. https://doi.org/10.1111/j.1464-0597.2008.00348.x