© 2019 American Psychological Association
0033-295X/19/$12.00

2019, Vol. 126, No. 6, 865–879
http://dx.doi.org/10.1037/rev0000156

The ‘Paradox’ of Converging Evidence

Clintin P. Davis-Stober
University of Missouri

Michel Regenwetter
University of Illinois at Urbana-Champaign

We explore the implication of viewing a psychological theory as the logical conjunction of all its predictions. Even if several predictions derived from a theory are descriptive of behavior in separate studies, the theory as a whole may fail to be descriptive of any single individual. We discuss what proportion of a population satisfies a theory’s joint predictions as a function of the true effect sizes and the proportion of variance attributable to individual differences. Unless there are no individual differences, even very well replicated effects may fail to establish that the combination of predictions that have been tested accurately describes even one person. Every additional study that contributes another effect, rather than strengthening support for the theory, may further limit its scope. Using four illustrative examples from cognitive and social psychology, we show how, in particular, small effect sizes dramatically limit the scope of psychological theories unless every small effect coincides with little to no individual differences. In some cases, this ‘paradox’ can be overcome by casting theories in such a way that they apply to *everyone* in a target population, *without exception*. Rather than relegating heterogeneity to a nuisance component of statistical models and data analysis, explicitly keeping track of heterogeneity in hypothetical constructs makes it possible to understand and quantify theoretical scope.

Keywords: effect size, inference, scientific reasoning fallacy, theoretical scope

This article shines a critical light on how a broad spectrum of research in psychology formulates, supports, and promotes many of its theories. Current meta-analyses and replication efforts within psychology attempt to provide a clear and comprehensive picture of the totality of evidence for or against a theoretical claim. Instead, we are chiefly concerned with theoretical scope: *What proportion of the population of interest satisfies a theoretical claim and what proportion is an exception to the theory?* This question is intimately connected to replication and what constitutes evidence for or against a theory. Despite the simplicity of this question, answering it is not a straightforward endeavor.

Rather than discuss methodological advances, we focus on theoretical scope. Replication is a very important methodological tool to establish whether a result is real. In many cases in psychology, the result under consideration is cast as an effect size. Hence, replication helps to evaluate whether a collection of effects exists and whether the estimated effect sizes are accurate. Likewise, meta-analysis is an important methodological tool to combine evidence across studies into an overall estimate of one or more effect sizes and to determine moderators that increase or reduce effect sizes. Throughout this article, we act as if we had completely accurate access to the true effect sizes. This is the best-case scenario of perfect reproducibility and perfect replicability: The scholar knows that each effect is true and knows exactly what each effect size is. Using simple quantitative arguments, we discuss the scope of psychological theories in the presence of genuine individual differences.

We spell out a ‘paradox’ of converging evidence: Most of us like to believe that, as we keep testing predictions from a theory in one study after another, the more predictions are empirically supported with good effect sizes, the better the theory has stood the test of time. However, this heuristic line of reasoning is logically

This article was published Online First August 15, 2019.

Clintin P. Davis-Stober, Department of Psychological Sciences, University of Missouri; Michel Regenwetter, Department of Psychology and Department of Political Science, University of Illinois at Urbana-Champaign.

This work was supported by National Science Foundation grants SES 14-59866 (PI: Clintin P. Davis-Stober) and SES 14-59699 (PI: Michel Regenwetter), as well as National Institutes of Health grant (K25AA024182, PI: Clintin P. Davis-Stober). Both authors have contributed equally to this project. Both authors have read and approved the final manuscript. The only “data” used in this article are published response proportions and Cohen’s d statistics from the literature. Therefore, this project is not subject to review by an Institutional Review Board. Any statements expressed in this publication are those of the authors and need not reflect the views of colleagues, funding agencies, or academic institutions. The authors are thankful to research assistants Ying Guo and Bryanna R. Fields for critical feedback, as well as for their assistance with a literature search, data retrieval, computations, and manuscript preparation. Yu Huang has also contributed to our insights. We thank Martin Schweinsberg for his help with access to the data from the pipeline project. Part of this work was presented to the 10-2-2017 *Student Faculty Seminar* in Political Science at the University of Illinois at Urbana-Champaign. Some of this work was also presented at the 2018 meeting of the Society for Mathematical Psychology. We thank the audience in both events for useful feedback.

Correspondence concerning this article should be addressed to Clintin P. Davis-Stober, Department of Psychological Sciences, University of Missouri, 219 McAlester Hall, Columbia, MO 65211. E-mail: stoberc@missouri.edu

invalid. In actuality, the more predictions we test, the more we learn about the limitations of the theory because every additional effect size potentially reduces the number of people in the population for whom the theory as a whole is descriptive. In particular, with every small effect size, the pool of potential "exceptions to the rule" becomes larger. While every additional effect reinforces the abstract notion of an "association" between an independent variable and a phenomenon, in contrast, additional effects play an opposite role when concretely thinking about numbers of individuals who satisfy a theory versus those who are exceptions to it.

It is common to test a psychological theory by deriving k many distinct predictions and then testing each prediction in a separate study or experiment. We build on Regenwetter and Robinson (2017, 2019) who showed that certain heuristic types of scientific reasoning in behavioral decision research are subject to logical reasoning fallacies. While Regenwetter and Robinson (2017, 2019) focused on behavioral decision research, this article expands the horizon to a broad range of psychology paradigms. We document far more generally that we commit a *fallacy of composition* when, as we establish that more and more parts of a theory are sometimes descriptive of some people's behavior, we become more and more confident that the theory as a whole is descriptive of someone's behavior. 'Paradoxically,' the opposite is true.

Thinking of a psychological theory as a "collection of predictions that each hold with exceptions" leads us down a rabbit hole of logical reasoning fallacies. Consider Newton's Law of Gravity. What if we knew for a fact that, while all other fruit drop to the ground when falling off a tree, a substantial fraction of lemons floated in midair instead? Would we consider those lemons that drop an example of weak, but nonetheless converging and supporting evidence in favor of Newton's Law of Gravity? The Law covers important differences having to do with masses of objects and the distances between them. It holds universally as long as the objects are not too small and not moving too fast. The domain of subatomic particles is not treated as giving weak but converging evidence supporting the Law of Gravity. It is treated as being outside the scope of the theory. Subatomic particles are exceptions to Newton's Law of Gravity. The natural sciences, as a matter of routine, study the scope of their theories. The purpose of this article is similar: Starting from the premise that psychological constructs are subject to genuine interindividual differences, we ask about the scope of a psychological theory. More specifically, we ask what proportion of a population is accurately described by a given psychological theory, based on the effect sizes of various predictions derived from that theory. Another way to state the same goal is to say that we aim to determine how big a proportion of the population is an exception to the theory.

We introduce our line of thinking with a simple illustration, using a very famous decision making study on framing effects by Tversky and Kahneman (1981). After that example, we move away from decision making to a broad range of cognitive and social psychology paradigms.

Motivating Example: Framing Effects

We revisit Tversky and Kahneman's (1981) discussion of framing effects, based on their Decision Problems 1–2 and 5–10. Problems 1–2 are two formulations of the famous Asian Disease problem, with Problem 1 stating a choice between two immunization strategies in terms of the number of lives saved, whereas Problem 2 frames the exact same strategies in terms of number of lives lost. To be consistent, a person who chooses A in Problem 1 must choose C in Problem 2 and a person who chooses B in Problem 1 must choose D in Problem 2. Tversky and Kahneman (1981) reported that 72% of their sample choose A over B, and that 22% of another sample choose C over D. If Tversky and Kahneman's measurement was perfect, namely if, indeed, exactly 72% of the population genuinely prefer A to B, and 22% of the population prefer C to D, then what can we say about the proportion of the population that either satisfy or violate classical models? The maximum proportion who can prefer A and C is the 22% who choose C in Problem 2. The maximum proportion of the population who can prefer both B and D is the 28% who chose B in Problem 1. In other words, at most 22% + 28% = 50% are consistent, hence at least half the population is an exception to classical theory and, instead, subject to this framing effect. At the other extreme, the largest possible proportion of the population who might prefer both A and D is the 72% who chose A in Problem 1, whereas the maximum proportion of the population who might prefer B and C are those 22% who chose C in Problem 2. In all, assuming that Tversky and Kahneman carried out perfectly accurate and externally valid measurements, anywhere between half and 94% of the population fall victim to the Asian Disease framing effect.

We now consider combinations of tasks. We skip Tversky and Kahneman's Problems 3 and 4 to avoid some distracting ambiguities. In their Problems 5, 6, and 7 a decision maker is consistent when either preferring A, C, and E or preferring B, D, and F. We treat any other combination as a framing effect. Responses to Decision Problems 8 and 9 are consistent if they match, that is, "Yes" (Y) in both cases or "No" (N) in both cases. The same applies to the two versions of Problem 10, where the decision maker needs to decide whether it is worth a 20 minute detour to save $5 off a $125 purchase or $5 off a $15 purchase. We denote the former as Problem 10L and the latter as Problem 10S. Consistency holds when a decision maker either responds Y to both questions or N to both questions.

Tables 1 and 2 show two hypothetical distributions of decision makers consistent with the data in Tversky and Kahneman (1981). In each table, the proportions of choices for each decision problem, given in the bottom two rows, match those reported by Tversky and Kahneman (1981).

In Table 1, altogether 50% of the hypothetical population is completely free of framing effects. It is also clear from our calculations above that this is the largest possible proportion of such people who could satisfy consistency across the board (i.e., regardless of how framing is operationalized): Tversky and Kahneman's hypothesis that *some* people fall victim to framing effects is supported by the insight that, taking their values at face value, at least 50% of the population must be exceptions to classical theory (consistency). The existence of framing effects is based on documenting that classical theory has limited scope. We can also reverse roles and ask about the scope of framing effects: How many people show framing effects systematically across tasks? In Table 1, 36% fall into that category (marked "F.E.").

Table 1
Hypothetical Distribution of Decision Makers Where Half of the Population is Completely Immune to Framing Effects Across Problems 1–2, 5–10 of Tversky and Kahneman (1981), 14% Show a Framing Effect (Boldface) in Some But Not All decisions, and 36% Show Framing Effects Everywhere (Marked F.E.)

	Decision problem and frame									Proportion of the population
	1	2	5	6	7	8	9	10L	10S	
Consistency	A	C	A	C	E	Y	Y	Y	Y	22%
	B	D	A	C	E	Y	Y	Y	Y	7%
	B	D	A	C	E	Y	Y	N	N	13%
	B	D	B	D	F	Y	Y	N	N	4%
	B	D	B	D	F	N	N	N	N	4%
Mixed	**A**	**D**	B	D	F	N	N	N	N	8%
	A	**D**	B	D	F	**Y**	**N**	N	N	3%
	A	**D**	B	D	F	**Y**	**N**	**N**	**Y**	3%
F.E.	**A**	**D**	**A**	**C**	**F**	**Y**	**N**	**N**	**Y**	32%
	A	**D**	**A**	**D**	**F**	**Y**	**N**	**N**	**Y**	4%
Tversky & Kahneman (1981)	72%	22%	78%	74%	42%	88%	46%	29%	68%	
	A	C	A	C	E	Y	Y	Y	Y	

Naturally, because Tversky and Kahneman establish the existence of framing effects by showing that many people must be exceptions to classical theory, then one can apply a similar question to framing: What sorts of exceptions to framing effects are consistent with Tversky and Kahneman’s data? Table 2 considers a hypothetical distribution in which, although everyone shows at least one framing effect, nobody shows it in more than two of the four problem sets that we consider. In other words, in this hypothetical distribution (whose marginal distribution matches Tversky and Kahneman’s data), nobody shows framing effects systematically across the four problem sets.

Taken together, Tables 1 and 2 highlight that, while Tversky and Kahneman (1981) have shown that *some people sometimes fall victim to framing effects*, they did not study the *scope* of this phenomenon. Their results leave open a broad range of possibilities as to what proportion of the population is subject to which framing effect and under what circumstance. Tversky and Kahneman (1981) followed a common path in psychology: They motivated their own theory by establishing that classical theory has insufficient scope. Yet, while showing that their theory (Cumulative Prospect Theory) can account for some phenomena that classical theory cannot, they did not, in turn, evaluate the scope of their own theory. In our view, psychology in general has much to gain from moving beyond this very common research strategy. Scholarly articles, as a matter of routine, should investigate and discuss their own scope.

This leads us to the purpose of the present article: The ‘paradox’ of converging evidence. While it seems on the surface that each additional study in Tversky and Kahneman (1981) provides additional evidence in favor of their proposed Cumulative Prospect Theory, two important insights are apparent from the constructions in Tables 1 and 2: First, with every additional study, the proportion of respondents who are consistent with classical theory can only decrease (or remain the same). In that sense, every additional study has the potential to add evidence that *some* people *sometimes* fall victim to *some* framing effects and violate classical decision theory. Hence, every additional study can add more evidence against the theory being challenged. On the flip side, however, with every additional study, the proportion of people who show framing across all studies can also only decrease (or remain the same at

Table 2
Hypothetical Distribution of Decision Makers Where Nobody Succumbs to All Framing Effects (Boldface) Across Problems 1–2, 5–10 of Tversky and Kahneman (1981)

	Decision problem and frame									Proportion of the population
	1	2	5	6	7	8	9	10L	10S	
Nobody shows F.E.s consistently	**A**	**D**	A	C	E	Y	Y	**N**	**Y**	39%
	A	**D**	A	C	E	Y	Y	Y	Y	3%
	A	**D**	**A**	**C**	**F**	Y	Y	N	N	4%
	A	**D**	**A**	**C**	**F**	N	N	N	N	4%
	B	D	**A**	**C**	**F**	N	N	N	N	8%
	B	D	**A**	**C**	**F**	**Y**	**N**	N	N	16%
	B	D	**A**	**D**	**F**	**Y**	**N**	Y	Y	4%
	A	C	B	D	F	**Y**	**N**	Y	Y	22%
Tversky & Kahneman (1981)	72%	22%	78%	74%	42%	88%	46%	29%	68%	
	A	C	A	C	E	Y	Y	Y	Y	

best). Just as every additional study further limits the scope of the classical theory, so does it also further delineate the scope of the new theory that is being proposed in its place. 'Paradoxically,' each additional study does not accumulate further evidence in favor of a proposed theory.

A common way to paraphrase Tversky and Kahneman (1981), is to claim that 'people' typically choose A (72%) and D (78%) in Problems 1 and 2; A (78%), C (74%) and F (58%) in Problems 5–7; Y (88%) and N (54%) in Problems 8 and 9; N (71%) and Y (68%) in Problems 10L and 10S. In the next section we will provide formulae that allow us to calculate upper and lower bounds on the number of people who have this specific preference pattern, taking Tversky and Kahneman's proportions at face value. The upper bound is .54 because no more than 54% answered No in Problem 9. According to Equation 1 in the next section, the lower bound is given by the following quantity:

$$\max\{1 - 9 + .72 + .78 + .78 + .74 + .58 + .88 + .54 + .71 + .68, 0\} = \max\{-1.59, 0\}.$$

At most 54% of people have that prototypical framing induced preference pattern ADACFYNNY, and at worst, nobody does. We have already seen an example of the latter in Table 2.

It is important to notice that the theory of framing risks becoming vacuous if cast as a logical disjunction of predictions (a person is subject to framing effects if she or he has framing dependent preference in at least one of the problem sets, that is, in Problems 1 and 2, OR in Problems 4, 5, and 6, OR in Problems 8 and 9, OR in Problems 10L and 10S). In our view, with a design like that in Tversky and Kahneman (1981), this claim becomes virtually immune to falsification. In this article, we conceptualize each theory as a logical conjunction of its predictions: If one claims that a treatment, say, enhances memory, then this is a claim regarding a hypothetical construct that, if true, must hold regardless of measurement design. To draw an analogy, suppose we have extremely high-quality photos of object E from different angles of view, all of them designed to give a clear view of the object, which we have hypothesized to be an elephant: In our opinion, to declare that E is indeed an elephant, it does not suffice that it appears like an elephant in one or two pictures. Rather, it ought to appear like an elephant from every 'reasonable' angle of view. By the same line of reasoning, a theory's predictions must hold jointly across all experimental conditions that 'properly' study the treatment effect. As we unpack these ideas, we will first assume measurement without error and then proceed to accommodate noisy measurement.

Basics

For the rest of the article, we consider the common setting in which treatment and control groups are being compared. In this section, we refer to the hypothetical constructs themselves, not to measurements. We will later expand our discussion to tackle the difficulties of compounding hypothetical constructs with measurement error. According to the theory at hand, the treatment i produces some *feature* F_i in the comparison between treatment group i and its control. We treat each feature as binary: A person either does or does not have it. A person has the feature F_i, say, if she genuinely succumbed to the Asian Disease framing effect, if she genuinely improved her memory, if she genuinely changed attitude, if she genuinely switched attention, if her cortisol level was genuinely higher. In particular, we cover the standard approach commonly found in psychology: A scholar derives several distinct predictions and evaluates each of these in a separate experiment or in a separate treatment manipulation. For now, when considering a change or difference in a quantitative variable, within person, we do not distinguish between a large and small difference, we only consider whether it occurs or not.

What proportion of the population is adequately described by a given psychological theory? In other words, how many people satisfy all the predictions made by the theory? Let p_i denote the proportion of the population that satisfies feature F_i, say, genuinely enhanced memory, or, say, a genuinely different attitude, and so forth, when undergoing treatment i. If we derive k many different predictions from a theory, those members of the population who are accurately described by the theory must, at the very least, satisfy all of the features $F_1, F_2, \ldots, F_k$. The proportion of the population that satisfies all of the features is, in turn, at best, $\min\{p_i, 1 \leq i \leq k\}$: If 55% of the population has feature F_1 and 60% of all people have feature F_2 then, at best, 55% have both features. What can we say about the scope of the theory? In other words, how many people, at least, in the population, satisfy all of the features $F_1, F_2, \ldots, F_k$? If $p_1 = .55$ and $p_2 = .6$, then the worst case scenario is that those 45% of the population who fail to satisfy feature F_1 are all included among those 60% of the population who satisfy feature F_2, in which case only 15% of the population would satisfy both features. In this worst case scenario, 85% of the population are exceptions to the theory. In general, for k many features, the proportion of people who satisfy all k features is bounded from below by

$$\max\left\{\min_j\{p_j - \sum_{i \neq j}(1 - p_i)\}, 0\right\} = \max\left\{1 - k + \sum_{\ell=1}^{k} p_\ell, 0\right\}. \tag{1}$$

In the above example when $k = 2$, $p_1 = .55$ and $p_2 = .6$, this becomes

$$\max\{1 - 2 + .55 + .6, 0\} = \max\{.15, 0\} = .15.$$

In this example, at worst 15% and at best 55% of the population satisfy both feature F_1 and feature F_2. Between 15% and 55% of the population satisfy both predictions of the theory. Therefore, unless a theory only predicts these two features and no others, the scope of the theory is at best between 15% and 55% of the population.

If we derive another, third, prediction from the theory and 75% of the population satisfies the predicted feature, that is, if $p_3 = .75$, then

$$\max\{1 - 3 + .55 + .6 + .75, 0\} = \max\{-.1, 0\} = 0.$$

With these three predictions and these proportions of the population who satisfy each, we conclude that, at best 55% of the population satisfy all three features, and, at worst, nobody does. The full theory, which is the conjunction of all its predictions not necessarily just these three, is descriptive of 55% of the population, at best. At worst, it is descriptive of *nobody*.

This is the 'paradox' of converging evidence: With every additional prediction, unless that prediction holds true for *everybody*, the support of the theory erodes rather than builds. In the worst case scenario, an extra prediction is descriptive only of people who violate at least one of the other predictions. Keep in mind that external validity requires a theory’s predictions to hold in numerous additional situations beyond those studied in the lab. If we consider the prediction that is descriptive of the fewest people and if we consider who it accurately describes, then in the *best case scenario*, those same people happen to also satisfy all the other features we have predicted from the theory. More specifically, a theory may be valid for a *target population* that may or may not include all people. If studying only members of the target population, then the predictions from this theory hold for everybody, hence there is no ‘paradox’ of converging evidence.

We now expand beyond hypothetical constructs to include measurement error. We consider the implications of these insights for the example of Cohen’s *d*, a very commonly used effect size in the behavioral sciences. Our illustration uses strong simplifying assumptions that serve the purpose of yielding easy closed-form formulae and concrete numbers so as to support a more general point. Our general conceptual argument about potentially accumulating exceptions to a theory rather than accumulating evidence in its favor does not depend on any particular effect size measure.

The ‘Paradox’ of Converging Evidence for Cohen’s *d* Effect Sizes

Let μ,μ' denote the population means of a dependent variable under the treatment and control conditions, respectively. Suppose that a theory[1] makes a prediction of the form $\mu > \mu'$. Suppose that we know the effect size, that is, we know the objective value of Cohen’s $d = (\mu - \mu')/\sigma$, where σ^2 is the variance of the dependent variable in either condition.

If an experimental design produces responses of the same person in both the treatment and the control condition, then we can compare those responses. In the Appendix, we show that if the dependent variable is normally distributed, and the two responses are stochastically independent, then the probability that a randomly picked respondent displays a larger value on the treatment response than in the control is $\Phi(d/\sqrt{2})$, where Φ denotes the cumulative distribution function of the standard normal. Many experiments do not actually collect responses to both experimental conditions from the same person. We are interested in theories that specify how a given person’s hypothetical construct depends on the experimental condition. For example, a theory may explain how the same person’s memory performance depends on the presence or absence of a given treatment.

Regardless of whether a person is exposed to both treatment and control, or just to one of the two, we assume that each person has a single, fixed (and unobservable) true value of the hypothetical construct in each condition. Suppose that the dependent variable is a sum of two independent normal variables: one captures individual differences in the value of the hypothetical construct, the other has mean zero and captures measurement noise. Therefore, the variance of the dependent variable can be written as the sum of two variances, the variance σ^2_{diff} associated with individual differences in the hypothetical construct and the variance σ^2_{err} associated with measurement error. Let

$$q = \frac{\sigma^2_{\text{diff}}}{\sigma^2} = \frac{\sigma^2_{\text{diff}}}{\sigma^2_{\text{err}} + \sigma^2_{\text{diff}}}$$

denote the proportion of the variance in the dependent variable that is associated with individual differences.[2] What is the probability that a randomly selected person’s true value of the hypothetical construct in the treatment condition is larger than the true value in the control condition? If there are no individual differences, that is, $\sigma^2_{\text{diff}} = 0$ and if $\mu > \mu'$, then everybody has the identical larger value μ of the hypothetical construct in the treatment condition. On the other hand, if $\sigma^2_{\text{diff}} \neq 0$, then, as the Appendix shows, the probability that a randomly selected person’s true value of the hypothetical construct in the treatment condition is larger than the true value in the control condition, is

$$\Phi(d/\sqrt{2q}). \tag{2}$$

Now, suppose that the theory predicts three effects, $\mu_i > \mu'_i$ for $i = 1, 2, 3$. Imagine that a researcher conducts three separate experiments on distinct samples of respondents from the same population. Suppose that, thanks to very high quality data, large sample size, and/or replication, the scholar has correctly identified the three true effect sizes: $d_1 = .4$, $d_2 = .7$ and $d_3 = .8$, one medium and two large effect sizes, in the terminology of Cohen (1988). What is the probability that a randomly selected respondent has higher levels of the hypothetical constructs in all three treatments? Because the theory presumably implies additional predictions besides those that are being tested, this is, at best, the proportion of people in the population of whom the theory is descriptive. This is an upper bound on the theory’s scope. We can answer this question if we know or assume the proportion q_i of the variance in the dependent variable for effect *i* that is attributable to individual differences in the value of the hypothetical construct. Suppose that $q_1 = q_2 = q_3 = .8$, that is 80% of the variance, in each dependent variable, is due to individual differences. Then,

$$p_1 = \Phi\left(\frac{.4}{\sqrt{2 \times .8}}\right) = .62; \quad p_2 = \Phi\left(\frac{.7}{\sqrt{2 \times .8}}\right) = .71;$$

$$p_3 = \Phi\left(\frac{.8}{\sqrt{2 \times .8}}\right) = .74.$$

As a consequence, according to Equation 1, the lower bound on the number of people in the population who satisfy all three predictions of the theory (at the hypothetical construct level) is

$$\max\{1 - 3 + .62 + .71 + .74, 0\} = \max\{.07, 0\}.$$

Between 7% and 62% of the population satisfy all three predictions.

If we suppose, instead, that $q_1 = q_2 = q_3 = .25$, that is, if only a quarter of the variance is due to individual differences, then,

[1] The Appendix provides formally precise statements of all the assumptions stated verbally here.

[2] Readers familiar with classical test theory may recognize the connection between *q* and the *reliability coefficient*, defined as the square of the correlation, in the population, between the observed values and the true values of the hypothetical construct (see, e.g., Webb, Shavelson, & Haertel, 2006). In test theory, reliability is used to quantify internal consistency of true scores in psychometric tests.

$$p_1 = \Phi\left(\frac{.4}{\sqrt{2 \times .25}}\right) = .71;$$

$$p_2 = \Phi\left(\frac{.7}{\sqrt{2 \times .25}}\right) = .84;$$

$$p_3 = \Phi\left(\frac{.8}{\sqrt{2 \times .25}}\right) = .87.$$

As a consequence, according to Equation 1, the lower bound on the number of people in the population who satisfy all three predictions of the theory is

$$\max\{1 - 3 + .71 + .84 + .87, 0\} = \max\{.42, 0\}.$$

Between 42% and 71% of the population satisfy all three predictions of the theory. Reducing the proportion of variance corresponding to individual differences has dramatically raised the lower bound.

Now, suppose in either case, that we derive one more prediction. Suppose that the objective effect size is small, say, $d_4 = .15$ and suppose that, in this experiment, 30% of the variance is due to individual differences. Then

$$p_4 = \Phi\left(\frac{.15}{\sqrt{2 \times .3}}\right) = .58.$$

As a consequence, according to Equation 1, the lower bound on the number of people in the population who satisfy all four predictions of the theory is

$$\max\{1 - 4 + .62 + .71 + .74 + .58, 0\} = \max\{-.35, 0\} = 0,$$

when $q_1 = q_2 = q_3 = .8$, and, respectively,

$$\max\{1 - 4 + .71 + .84 + .87 + .58, 0\} = \max\{0, 0\} = 0,$$

when $q_1 = q_2 = q_3 = .25$. For either case, at best 58% of the population satisfies all four predictions. At worst, *nobody* satisfies them, so the theory is descriptive of *nobody*. Despite two large effects, one medium, and one small effect, all of which are known to be genuine effects with accurate effect sizes, we have failed to establish that the theory is descriptive of *anybody*. Furthermore, the additional small effect $d_4 = 0.15$, rather than lending a little bit of additional support to the theory, places a major constraint on the possible scope of the theory.

In the example, we have assumed that everybody has, in each condition, a fixed value of the hypothetical construct. Together with the normal distribution assumption for individual differences with positive variance, this means that not everybody can satisfy any given prediction of the theory (e.g., genuinely enhanced memory in the treatment condition). The 'paradox' of converging evidence results from the fact that when each prediction only holds for part, but not all, of the population, then each prediction has the potential of further eroding the evidence that anyone is accurately described by the full theory as a whole. In other words, if a small effect is simply a "prediction with many exceptions," then we need to worry about the degree to which exceptions accumulate across multiple predictions. This problem is particularly serious if an additional effect is small and if a large portion of the variance in responses is due to individual differences. These insights raise a very serious question about external validity. Suppose that, thanks to successful replication and large scale meta-analyses, we have objective knowledge of a few effect sizes, some of which are small. If the scope of the theory, limited to that handful of treatments in the laboratory, is very small, what can one extrapolate about its scope for 'stimuli' and 'treatments' outside the laboratory?

For the remainder of the article, we revisit a few published studies and consider their findings through the lens of Equations 1 and 2. For simplicity, we concentrate on studies that used three or more Cohen's *d* effect sizes to claim support for their theory.

Illustrative Examples

Revisiting Leavitt and Christenfeld (*Psychological Science*; Leavitt & Christenfeld, 2011)

Leavitt and Christenfeld (2011) considered the effect of story spoilers on reader enjoyment and reported that (p. 1153) "giving away [. . .] surprises makes readers like stories better" regardless of whether the story involved a "twist at the end," "solved the crime," or was "poetic." "In all these types of stories, spoilers may allow readers to organize developments, anticipate the implications of events, and resolve ambiguities that occur in the course of reading." Specifically, according to Leavitt and Christenfeld (2011),

> Subjects significantly preferred spoiled over unspoiled stories in the case of both the ironic-twist stories (6.20 vs. 5.79), p =.013, Cohen's $d = 0.18$, and the mysteries (7.29 vs. 6.60), p =.001, $d = 0.34$. The evocative stories were appreciated less overall, [. . .], but subjects again significantly preferred spoiled over unspoiled versions (5.50 vs. 5.03), p =.019, $d = 0.22$ (pp. 1153–1154).

The message of the article is one of consistent and mutually reinforcing evidence for a general theoretical claim. In our view, the theoretical claim is clearly a logical conjunction: "Spoilers don't spoil stories" (regardless of genre).

It seems to us natural to question that claim and, instead, conjecture that there are individual differences at play here: Presumably some people genuinely prefer spoilers (and to different degrees) and some genuinely prefer unspoiled stories.[3] This is a good case study for investigating the proportion of the population that genuinely satisfies the theoretical claim. We consider what conclusions one could draw in the idealized event that extensive meta-analyses and/or replication studies successfully reinforced the reported effect sizes. We go so far as to suppose that Leavitt and Christenfeld (2011) determined objective and accurate effect sizes. Taking $d_1 = 0.18$, $d_2 = 0.34$, and $d_3 = 0.22$, at face value, we can ask what proportion of the population genuinely likes stories better with a spoiler, depending on how much variance is attributable to error or individual differences, respectively. For simplicity, we further assume that the proportion of variance due to individual differences, q, is the same across the three experiments.

Figure 1 shows the upper and lower bounds on the proportion of the population who genuinely prefer spoilers, as a function of q, using the combination of Equations 1 and 2. Moving from left to right, the graphs show the upper and lower bounds on the theory's scope for one ($d_1 = 0.18$), two ($d_1 = 0.18$, $d_2 = 0.34$,), and three effect sizes ($d_1 = 0.18$, $d_2 = 0.34$, $d_3 = 0.22$). With only a single effect size (left most graph), the upper and lower

[3] Incorporating the notion that a given reader may also waiver or vary in this propensity would further complicate the interpretability of the results.

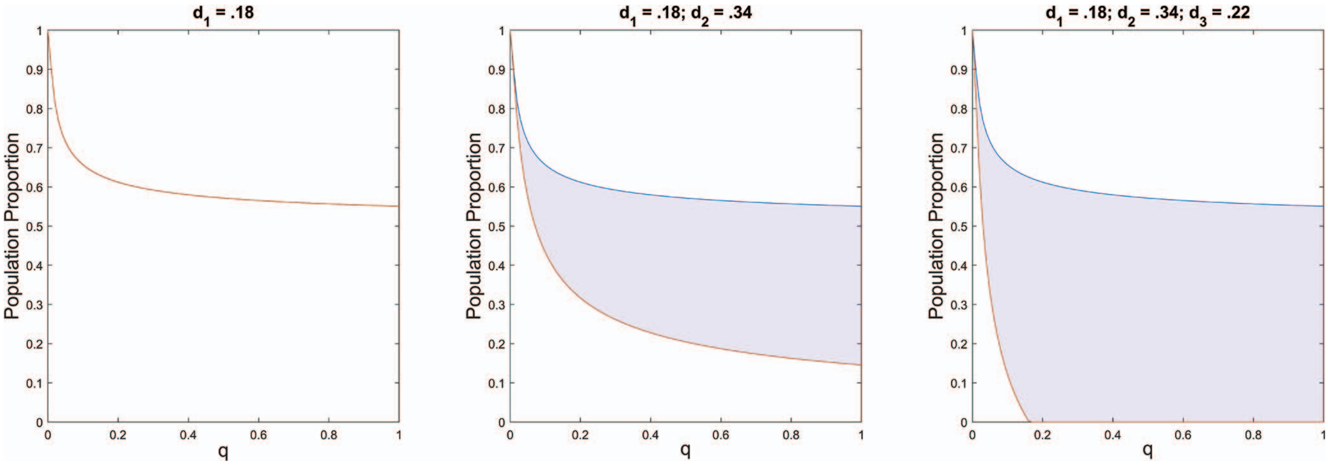

Figure 1. Upper and lower bounds on the proportion of the population who genuinely prefer spoilers in all three types of stories, as a function of the proportion q of variance due to individual differences (based on three effect sizes: $d_1 = .18$; $d_2 = .34$; $d_3 = .22$). See the online article for the color version of this figure.

bound are the same. For almost the entire range of q, we infer that roughly $\frac{1}{2}$ to $\frac{2}{3}$ of the population genuinely prefer spoilers in “ironic-twist” stories. Because we considered the lowest effect size first, the upper bound is the same across all three graphs. However, with every additional effect size, the lower bound decreases. In the right most graph, the lower bound exceeds zero only when less than about 16% of the variance is due to individual differences.

In our view, it goes without saying that some people sometimes appreciate spoilers. It also seems self-evident to us that the preference for spoilers cannot be a universal law that, in the way that Newton’s Law of Gravity applies to all fruit dropping from trees, would likewise apply exactly the same way to every reader. To us the main scientific value of this line of reasoning is that it permits one to delineate the scope of the theoretical claim. Taking the three effects sizes at face value, unless only a small portion of the variance is due to individual differences, Figure 1 shows that the proportion of people for whom spoilers genuinely do not spoil stories ranges anywhere from nobody to $\frac{2}{3}$ of the population at best. Should additional studies derive additional predictions from the theory and find even smaller effect sizes, then the upper bound would drop further, and the lower bound might become zero even for extremely small values of q.

In all, in our view, it would be a logical fallacy of composition to conclude, as the original article appears to state in its title, that irrespective of reader and genre, “Story Spoilers Don’t Spoil Stories.” Not only would it likely be a major mistake for book publishers and authors to focus exclusively on stories with spoilers, we also see little value in such sweeping generalizations for scientific discourse. In our view, the evidence presented in Leavitt and Christenfeld (2011), with the effect sizes taken at face value, allows for the possibility that everyone may be an exception to the theory. Future research on this paradigm could make individual differences part of the theory. This future research paradigm would need to take care not to succumb to another ‘paradox’ of converging evidence. For example, determining moderators that make the effects larger or smaller, per se, would not eliminate the ‘paradox’ of converging evidence, unless certain values of the moderator completely eliminated individual differences. One solution would be to incorporate heterogeneity into the theory in such a way that every prediction applies to everyone within a well-specified target population. We discuss this idea in more detail later.

Revisiting Foroughi et al. (*Journal of Experimental Psychology: General*; Foroughi, Werner, Barragán, & Boehm-Davis, 2015)

Foroughi et al. (2015) considered cognitive research on working memory and reading comprehension. They proposed a theory as a competitor and contrast to an earlier theory of Ericsson and Kintsch (1995). Their core theoretical claim was “that the transient portion of working memory is necessary for text comprehension” (p. 708). Foroughi et al. (2015) supported their claim by investigating whether (pp. 707–708) “interruptions while reading disrupt reading comprehension when the questions assessing comprehension require participants to connect and synthesize information across the passage.” They stated (p. 708): “In Experiment 1, we found that interruptions disrupted reading comprehension” a conclusion which they supported with a Cohen’s $d_1 = .85$ in a task that involved only comprehension questions. “In Experiment 2, we made the distinction between reading comprehension and recognition. Consistent with Experiment 1, we found that interruptions disrupted reading *comprehension*” a conclusion that they supported with a Cohen’s $d_2 = .95$ in a task that combined some comprehension questions with some recognition questions. “In Experiment 3, we replicated the findings from Experiments 1 and 2 and also found that adding a 15-s time-out period prior to each interruption prevented the disruption caused by interruptions when answering questions assessing comprehension” a conclusion that they supported with Cohen’s $d_3 = .70$ and $d_4 = .73$ in a task like that in Experiment 2 with additional time-out condi-

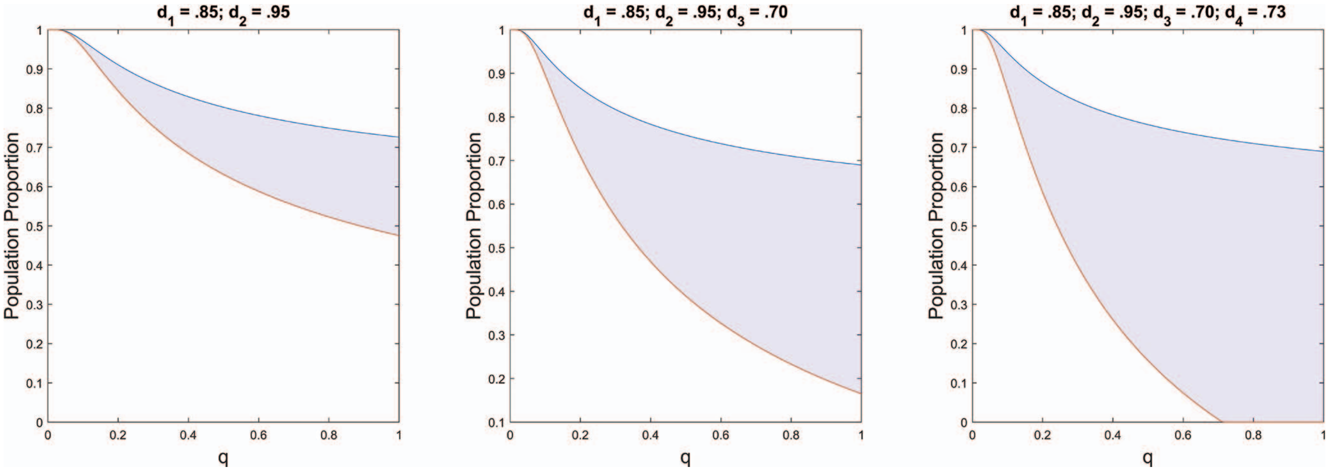

Figure 2. Upper and lower bounds, as a function of q, for the proportion of the population for whom the four predictions of Foroughi et al. (2015) are jointly descriptive (based on four effect sizes: $d_1 = .85$; $d_2 = .95$; $d_3 = .70$; $d_4 = .73$). See the online article for the color version of this figure.

tions mixed into the task, analyzed separately without, respectively with, the time-out. Once again, we take the effect sizes at face value.[4] In our view, this study very much resembles our analogy about photos of an object E that we have hypothesized to be an elephant. Each experimental condition is like getting yet another good angle of view of the *same* object. Therefore, just like every picture ought to reveal an elephant, each experimental condition ought to reveal that interruptions disrupt reading comprehension.

The graphs of Figure 2 show the upper and lower bounds as we move from two to three and eventually four effect sizes. Because Cohen's d for Experiment 1 is smaller than that for Experiment 2, the single curve we would get from just Experiment 1 is the same as the upper bound in the left most graph in Figure 2. Once again, as one incorporates one experiment after the other, rather than accumulating evidence "that the transient portion of working memory is necessary for text comprehension," one actually tempers that claim: As we move from the left graph to the right, the upper and lower bounds on the proportion of people in the population who satisfy the theoretical claim both decrease. While the overall picture is far less dire than that in Figure 1, especially if one considers a memory paradigm like this as less prone to individual differences, one or two additional studies with small effect sizes could still strongly reduce support for this theory as a theory of how "people's" memory works. There would be too many documented exceptions in order for the theory to hold at a high level of generality.

Because of the experimental design specifics in Foroughi et al. (2015), a case could be made that the first three Cohen's d values are replications of one and the same effect. Assuming just two effect sizes[5] $d_{avg} = \frac{.95+.85+.70}{3} = .83$ and $d_4 = .73$ gives the bounds in Figure 3.

Revisiting Armor et al. (*Psychological Science*; Armor, Massey, & Sackett [2008])

Armor, Massey, and Sackett (2008) considered "Prescribed Optimism" and asked "Is it right to be wrong about the future?" They suggested that personal predictions are often optimistically biased, not because of irrationality but rather because such behavior is consistent with people's values and beliefs. The theoretical claim was (p. 330) that people "believe optimistically biased predictions are ideal." The authors supported their claim with four Cohen's d effect sizes based on four predictions: They reported that people prescribe optimism bias ($d_1 = 0.93$); that people describe others ($d_2 = 0.80$) as well as themselves ($d_3 = 0.56$) as optimistically biased, and that people believe one should be even more optimistically biased ($d_4 = 0.24$). This is another case where big individual differences seem to us rather plausible. Figure 4 shows the corresponding upper and lower bounds on the proportion of the population for whom the theory of Prescribed Optimism is descriptive, assuming again that the reported effects exist and that their effect sizes are fully accurate.

The graph at the top left of Figure 4 shows the bounds for the single effect size $d_1 = 0.93$. Regardless of the value of q, the vast majority of the population appears to genuinely exhibit the first effect. The remaining graphs incorporate the additional information gained by the additional effect sizes associated with the additional predictions from the theory ($d_2 = 0.80$, $d_3 = 0.56$, $d_4 = 0.24$). Combining all four effects and assuming perfectly accurate effect sizes, one needs q below about 0.055 to obtain an upper bound above 70% of the population. If half the variance or more is due to individual differences, then we cannot conclude that even a single member of this population is accurately described by the "Prescribed Optimism" claim. The four Cohen's d values simply permit too many exceptions to the theory. A replication[6] of the first effect by the Open Science Collaboration (2015) estimated d slightly higher at $d_1 = 1.18$. Replacing $d_1 = 0.93$ by $d_1 = 1.18$

[4] We use the values from "Foroughi, Werner, Barragán, and Boehm-Davis, 2016."

[5] We average the first three d values because they were based on identical sample sizes and similar *SE*s and, once again, we take the resulting d at face value.

[6] See van t Veer, Lassetter, Brandt, and Mehta (n.d.).

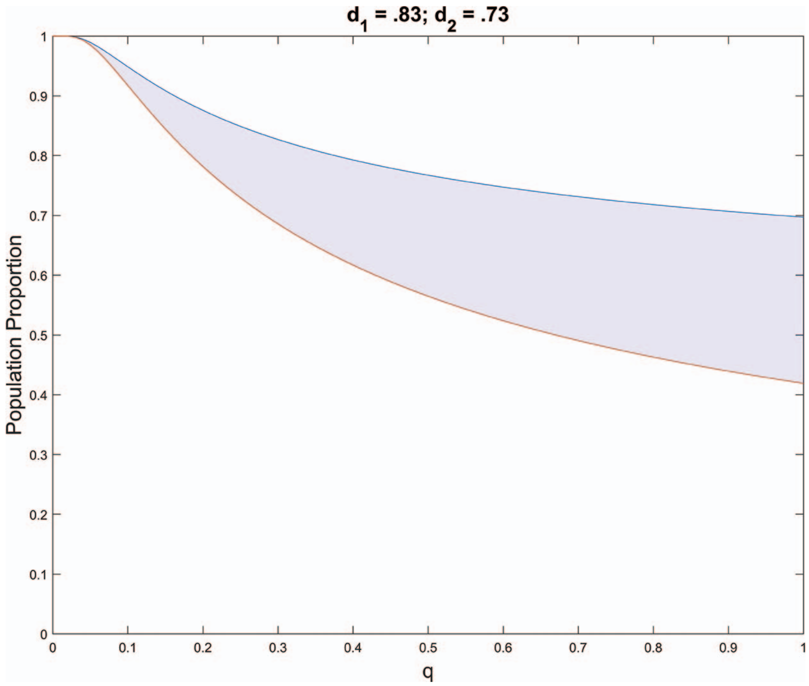

Figure 3. Upper and lower bounds, as a function of q, for the proportion of the population for whom the two predictions of Foroughi et al. (2015) are jointly descriptive (based on two effect sizes: d_{avg} = .83; d_4 = .73). See the online article for the color version of this figure.

only slightly changes the bounds we computed and does not affect our substantive conclusion.

Revisiting the Pipeline Project (*Journal of Experimental Social Psychology*; Schweinsberg et al., 2016)

As our last illustration, we consider Studies 1–6 of a large scale scientific crowdsourcing *pipeline project*, that is, “Pre-publication independent replications of a single laboratory’s research pipeline,” involving 25 laboratories from around the world (Schweinsberg et al., 2016). The six effects we consider are the (a) *bigot-misanthrope effect*, (b) *cold-hearted prosociality effect*, (c) *bad tipper effect*, (d) *belief-act inconsistency effect*, (e) *moral inversion effect*, and (f) *moral cliff effect*, respectively.

The theoretical thread linking these six effects comes from the *Person-Centered Account of Moral Judgment* (Uhlmann, Pizarro, & Diermeier, 2015), according to which “moral evaluations are frequently driven by informational value regarding personal character rather than the harmfulness and blameworthiness of acts” (Schweinsberg et al., 2016, p. 57). The gist of the claim is “that internal factors, such as intentions, can be weighed more heavily in moral judgments than objective external consequences” (Schweinsberg et al., 2016, p. 57). Uhlmann et al. (2015) clearly consider their theory to be strongly supported by an accumulation of evidence. They state (p. 73) that

> there is growing evidence that when it comes to moral judgment, human beings appear to be best characterized not as intuitive deontologists or consequentialists but as intuitive virtue theorists: individuals who view acts as a rich set of signals about the moral qualities of an agent and not as the endpoint of moral judgment.

This claim of external validity is hard to understand unless we treat the theory as a logical conjunction of its predictions.

Uhlmann et al. (2015) and Schweinsberg et al. (2016) suggest that, thanks to the accumulation of effects, as well as the successful replication of these effects, there is converging evidence in support of the Person-Centered Account of Moral Judgment. In our view, the framing of the theory makes it clear from the onset that this is intended, not as a universal theory of everyone or of every moral judgment, but rather as a theory that allows for potentially many exceptions. In particular, it makes sense to us to expect extensive individual differences, that is, to consider ‘large’ values of q. Hence, one should ask how the heuristic accumulation of evidence in support of the theory, as it is spelled out rhetorically in Uhlmann et al. (2015) for example, compares with the quantitative accumulation of possible exceptions to the theory that follows mathematically from the same effects, the same effect sizes, and the same studies.

We consider six meta-analytic replication estimates from the pipeline project (d_1 = 1.27; d_2 = 2.05; d_3 = .57; d_4 = .41; d_5 = .51; d_6 = .70). The huge scale and hybrid nature of the pipeline project provides compelling support for the existence and approximate size of the effects. Taking these well-estimated six effect sizes at face value, we generate upper and lower bounds on the proportion of the population of whom the Person-Centered Account of Moral Judgment is descriptive, as a function of the proportion of variance due to individual differences.

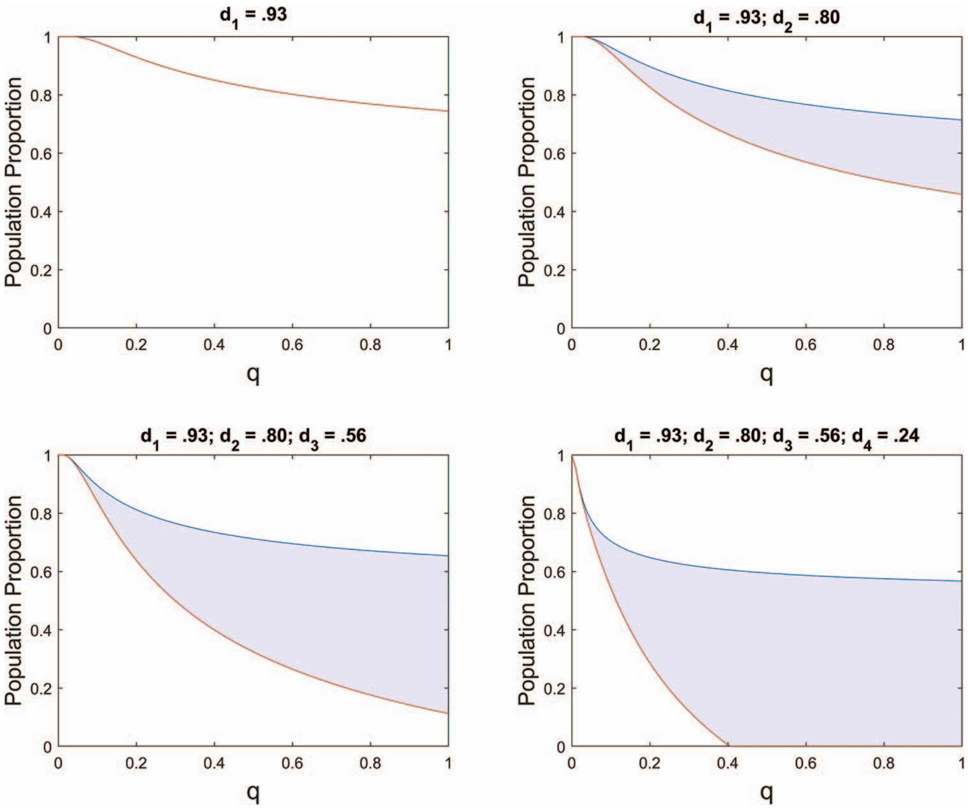

Figure 4. Upper and lower bounds, as a function of q, for the proportion of the population for whom the four predictions of Armor et al. (2008), derived from the Prescribed Optimism theory, are jointly descriptive. See the online article for the color version of this figure.

Figure 5 displays the bounds, as a function of q, for all six effects. For us to be certain that these six predictions are jointly descriptive of some people's behavior, no more than 30% of the overall variance can be attributable to individual differences. If only 10% of the variance is attributable to individual differences then we can be reasonably certain that these six predictions jointly hold for a majority of individuals in the population.

'Best Case Scenarios'

We have mentioned the possibility that, as a best case scenario, a theory might apply to *everyone* in some subpopulation. If a proportion p of the population satisfies all k distinct predictions of the theory then the proportion p_i of people satisfying prediction i must be at least p. Letting z_p denote the z-score such that $\Phi(z_p) = p$, this implies that, for the *ith* prediction,

$$p \leq p_i = \Phi\left(d_i/\sqrt{2q_i}\right), \quad \text{hence,}$$

$$q_i \leq \frac{d_i^2}{2 \times z_p^2}. \tag{3}$$

Because the effect size is squared in the numerator, this constraint implies that, for each prediction with a small effect size, the proportion of the variance due to individual differences would have to be tiny. Once again, throughout this section, we take the effect sizes at face value.

Table 3 shows the constraints derived in this fashion for the study of Leavitt and Christenfeld (2011) about story spoilers. The table shows what follows under the assumption that there is a subpopulation within which the theory holds *without exception*. For example, if there is a subpopulation that makes up 70% of the total population, and in this subpopulation spoilers genuinely do not spoil stories, regardless of genre, then, with the effects sizes taken at face value, we would need $q_1 \leq 0.059$, $q_2 \leq 0.21$, and $q_3 \leq 0.088$. Hence, if the theory applies, without exception, to everyone belonging to a subpopulation of that size, then the proportion of variance due to individual differences cannot be more than about 6% (ironic-twist), 21% (mystery), and 9% (poetic), respectively. If the theoretical claim that spoilers do not spoil stories is accurate for a subpopulation making up 90% of the population, taking the effect sizes at face value, then the proportion of variance due to individual differences cannot be more than about 1% (ironic-twist), 4% (mystery), and 2% (poetic), respectively.

Tables 4 and 5 show the corresponding best case scenarios for the memory theory of Foroughi et al. (2015) and for the Prescribed Optimism theory of Armor et al. (2008), respectively. Under the assumption that a subpopulation spanning 60% of the population satisfies the theory of Foroughi et al. (2015) *without exception*, we find that the proportion of variance attributed to individual differences is unconstrained. Under the assumption that a subpopulation

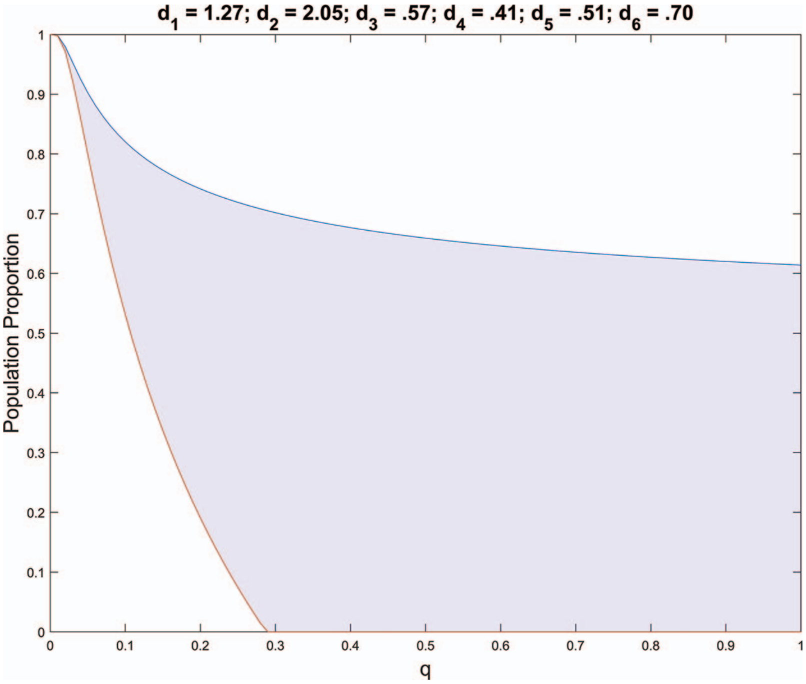

Figure 5. Upper and lower bounds, as a function of q, for the proportion of the population for whom the six predictions of Schweinsberg et al. (2016), derived from the Person-Centered Account of Moral Judgment, are jointly descriptive. See the online article for the color version of this figure.

spanning 90% of all people satisfies the theory without exception, we find that the percentage of variance attributable to individual differences must be less than 15% to 25%, depending on the effect size. The situation is different for the Prescribed Optimism theory of Armor et al. (2008). For example, under the assumption that a subpopulation spanning 80% of all people satisfies the theory *without exception*, the proportion of variance attributable to individual differences for $d = 0.24$ is at most 4%. This best case scenario analysis highlights once again that the scope of the theory is very limited unless one assumes that individual differences are minimal to nonexistent. To us, this seems rather implausible for a theory like Prescribed Optimism.

Table 6 illustrates a general point about the best case scenario, where those people who satisfy the first prediction of the theory are the exact same people who also satisfy the second prediction, and the third, and so forth; as long as effect sizes are large, it is possible that a large proportion of the population is well described by all predictions, even with a large proportion of the variance in the hypothetical construct due to individual differences. For consistently large effects across the board, the best case scenario offers a very optimistic and favorable outlook on a theory’s potential scope. However, as soon as we encounter predictions with modest or small effect sizes, such as those beyond the first two lines in Table 4, the situation becomes far more nuanced: Now, even in the best case scenario where there is maximal overlap between people who satisfy the various predictions, the only way that these predictions can hold jointly (without exception) for a substantially sized subpopulation is if individual differences are minimal to nonexistent for those predictions that yield small effect sizes.

Table 3

Best Case Scenario Where Spoilers Genuinely Don’t Spoil Stories (Regardless of Genre) for Everyone in a Subpopulation That Makes Up a Proportion p of the Population

	p	.6	.7	.8	.9	.95	.99
d	z_p	.2533	.5244	.8416	1.2816	1.6449	2.3263
.18	q_1	≤.25	≤.059	≤.023	≤.0099	≤.0060	≤.0030
.34	q_2	≤.90	≤.21	≤.082	≤.035	≤.021	≤.011
.22	q_3	≤.38	≤.088	≤.034	≤.015	≤.0089	≤.0045

Note. Each of the bottom three rows contains the effect size d and upper bounds on q_i, the proportion of variance attributed to individual differences, for six different values of p.

Table 4

Best Case Scenario for Foroughi et al.’s Memory Theory Under the Assumption That Everyone in a Subpopulation of Size p Satisfies the Theory Without Exception

	p	.6	.7	.8	.9	.95	.99
d	z_p	.2533	.5244	.8416	1.2816	1.6449	2.3263
.85	q_1	≤1	≤1	≤.51	≤.22	≤.13	≤.067
.9	q_2	≤1	≤1	≤.57	≤.25	≤.15	≤.075
.7	q_3	≤1	≤.89	≤.35	≤.15	≤.09	≤.045
.73	q_4	≤1	≤.97	≤.38	≤.16	≤.10	≤.049

Note. Each of the bottom four rows contains the effect size d and upper bounds on q_i, the proportion of variance attributed to individual differences, for six different values of p.

Table 5
Best Case Scenario for the Prescribed Optimism Theory Under the Assumption That Everyone in a Subpopulation of Size p Satisfies the Theory Without Exception

	p	.6	.7	.8	.9	.95	.99
d	z_p	.2533	.5244	.8416	1.2816	1.6449	2.3263
1.15	q_1	≤1	≤1	≤.93	≤.40	≤.24	≤.12
.80	q_2	≤1	≤1	≤.45	≤.19	≤.12	≤.059
.56	q_3	≤1	≤.57	≤.22	≤.10	≤.06	≤.029
.24	q_4	≤.45	≤.10	≤.041	≤.018	≤.011	≤.0053

Note. Each of the bottom four rows contains the effect size d and upper bounds on q_i, the proportion of variance attributed to individual differences, for five different values of p.

These concerns tie in with Smith and Little (2018) who warned that many journals now overvalue large sample sizes in the interest of detecting small effects. They highlighted the disconnects that emerge when evaluating theories about individual behavior with large sample between-subjects designs. Instead, Smith and Little (2018) advocated for "small-N" designs that test such theories with large amounts of information gathered from each of a few participants. Connecting Smith and Little's ideas with ours, if scholars collect large samples because they are looking for small effects, then they need to be aware that large individual differences may severely limit theoretical scope. On the other hand, large individual differences may also endanger the efficacy of "small-N" designs.

Solving for the various q_i values, as we have done in the above tables, also suggests that there are ways to evaluate theory scope empirically. For any given level of p, the tables provide a basis for possible hypothesis tests in studies that collect a suitable amount of data of the right kind, such as within subjects repetitions of the same measurements or within subject manipulations of experimental conditions. We leave the ensuing methodological questions and implications for other work. Our tables, together with the earlier figures, provide theoretical guidance and document the urgent need of the discipline to consider theory scope as a core issue that reaches beyond replicability and precision of effect size measurement.

Conclusion and Discussion

The standard rhetorical tool of citing various studies that report effects consistent with a given theory as converging evidence in favor of that theory is a reasoning heuristic with little logical merit. Finding, replicating, and accumulating nonzero effect sizes, and establishing that they are accurately measured does not accumulate evidence that the theory as a whole is descriptive of anyone. 'Paradoxically,' the contrary is true. This 'paradox' of converging evidence presents a wide-ranging challenge in the way that much of psychology formulates and tests theory. We have documented logical gaps in how psychologists often state theories, as well as how they often interpret their findings: Using Cohen's d effect sizes as a concrete and simple platform, we have laid out the fundamental disconnect between standard evidence 'accumulation' in the presence of several 'good' Cohen's d values, and the ability to infer whether a theory accurately describes any individual members of a population.

Replications and meta-analyses, while crucially important for determining whether effects are real and how large their effect sizes are, do not, per se, solve the 'paradox' of converging evidence. While scholars readily acknowledge the existence and importance of individual differences, there appears to be little awareness of the degree to which heterogeneity at the level of hypothetical constructs can erode the scope of a theory. The interaction between individual differences on the one hand and the accumulation of significant effects on the other hand, has a debilitating impact on the overall evidence supporting a theory. Rather than incorrectly considering well-established and well-estimated effects as converging evidence in support of a theory, psychologists should pay more careful attention to the potential accumulation of exceptions to their theories that result from a collection of effects, and from small effect sizes, in particular. Successful replication and successful meta-analyses only mean that we have a very good grasp of the effect sizes in question. The real challenge to psychological theory and to psychological science is the following: If everyone or nearly everyone is an exception to a given theory, then, we would argue, the theory is really not viable. A common item of wisdom in psychology is that "every theory is wrong." In our view, this is not a reason to be either cynical or nonchalant about the quality of psychological theories. Rather, evaluating upper and lower bounds on how many people are described jointly by a set of predictions from a theory provides a concrete and precise way to conceptualize and quantify the theory's scope.

Our approach may appear overly strict in that we treat a theory as the logical conjunction of its predictions. At the opposite extreme, a disjunction of predictions, to us, seems essentially non falsifiable. Besides evaluating scope, scholars who do not view their theory as a conjunction of its predictions ought to clarify how they conceive of theoretical scope. For example, they should spell out what predictions must hold jointly and how all predictions relate to each other.

We have used Cohen's d, under a simple experimental design, to document a specific version of the 'paradox' of converging evidence. More complex experimental designs would, in and of themselves, do little to resolve the problem. For analysis of variance (ANOVA) designs, for instance, each prediction would state a collection of main effects and interaction terms. Above, we have

Table 6
Best Case Scenario for the Person-Centered Account of Moral Judgment Under the Assumption That Everyone in a Subpopulation of Size p Satisfies the Theory Without Exception

	p	.6	.7	.8	.9	.95	.99
d	z_p	.2533	.5244	.8416	1.2816	1.6449	2.3263
1.27	q_1	≤1	≤1	≤1	≤.49	≤.30	≤.15
2.05	q_2	≤1	≤1	≤1	≤1	≤.78	≤.39
.57	q_3	≤1	≤.59	≤.23	≤.10	≤.06	≤.030
.41	q_4	≤1	≤.31	≤.12	≤.05	≤.031	≤.016
.51	q_5	≤1	≤.47	≤.18	≤.079	≤.048	≤.024
.70	q_6	≤1	≤.89	≤.35	≤.15	≤.091	≤.045

Note. Each of the bottom six rows contains the effect size d (from the pipeline project meta-analysis) and upper bounds on q_i, the proportion of variance attributed to individual differences, for six different values of p.

seen that $\mu > \mu'$ still allows some individuals to have smaller values of the hypothetical construct in the treatment group (e.g., even when memory is enhanced on average, some individuals may have genuinely impaired memory under the treatment). Likewise, there is little reason to assume that every individual’s values of the hypothetical constructs satisfy the same orders and dependencies as the ANOVA states for population means (e.g., even though, on average, memory is enhanced and attention is more focused in a fashion that depends on what is being remembered, some individuals may genuinely remember less, or some individuals’ attention may not depend on memory in the same way as that dependency holds on average). In summary, in the presence of individual differences, each ANOVA may support a theory’s prediction only for a portion of the population, hence leading to the ‘paradox’ of converging evidence in similar ways as we have documented for comparisons of two population means and Cohen’s *d*.

The most compelling way out of the ‘paradox’ is to make predictions that are predictive of *all* people. One possible solution is to reduce the scope and identify a suitably delineated target population, within which *everyone* satisfies each prediction. An example would be if, while allowing for measurement error, *all* people diagnosed with a certain disorder, had *all* predicted features, *without exceptions*. Haaf and Rouder (2017) developed a Bayesian framework for evaluating whether all individuals in a sample display an experimental effect, up to measurement error. Such frameworks would be useful in estimating the level of *q*. In the ‘Best Case Scenarios’ section, we have explored constraints we could derive from assuming that a subpopulation of relative size *p* satisfies a given theory *without exception*. In many cases, unless we allow that subpopulation to be rather small, we find strong constraints on *q*. This means that one would have to attribute the vast majority of variance in the dependent variable to erroneous measurement and not to individual differences in the hypothetical construct. In our view, scholars should pay great attention to the possible sources of variation and aim to measure them directly. Especially for small effect sizes, the only way that a theory can apply to a large portion of the population is to assume that individual differences are minimal or nonexistent.

Another possible solution to the ‘paradox,’ besides specifying a theory that applies only to a subpopulation, is to let the features $F_1, F_2, \ldots F_k$ be defined as probability statements that apply to everyone without exceptions: Instead of claiming, for example, that *k* different treatments each decrease anxiety for some but not all people, at the cost of making a collection of predictions that may jointly hold for nobody, a theory may rather predict that, for each individual, each treatment has probability exceeding, say, .80 of decreasing the person’s anxiety. It is important to distinguish between a claim that at least 80% of people have lower anxiety (that remains open to the ‘paradox’ across multiple such treatments) and a claim according to which every individual has a probability of at least .80 of experiencing lower anxiety. This type of distinction is empirically testable in principle, for example, by running a treatment repeatedly on the same person and/or by testing all predictions jointly on each individual.

Our general argument is intended to encourage a more comprehensive and more concrete approach to theory specification and evaluation that aims to go beyond “associations” between variables of interest. Short of knowing objective effect sizes, determining a theory’s scope needs to work hand-in-hand with meta-analyses and replication efforts. Methodologists are developing ever better tools to estimate effect sizes accurately. These include meta-analytic methods to mitigate the effects of publication bias (e.g., Guan & Vandekerckhove, 2016; van Assen, van Aert, & Wicherts, 2015) and improve reported effect sizes for individual articles (e.g., McShane & Böckenholt, 2017). The best meta-analyses consider subpopulations and other potential moderators, so as to get a better handle on what leads to large or small effects. For example, in a comprehensive meta-analysis of the effect of acute stress on episodic memory, Shields, Sazma, McCullough, and Yonelinas (2017) identified that the effects of stress on memory were less pronounced for women taking hormonal contraceptives. However, the effect of moderators on effect sizes, per se, does not establish theoretical scope. For example, a smaller effect size need not equate fewer individuals satisfying the effect of interest. Consider the women taking hormonal contraceptives. If the contraceptives reduced individual differences in episodic memory then the smaller effect could conceivably go hand-in-hand with more women actually showing memory deficits under acute stress. The smaller effect size for this subpopulation could come about if either more or fewer of these women were impacted, depending on the degree of individual differences in the hypothetical construct in that subpopulation.

Our argument is that the line of empirical inquiry cannot end with the establishment of effects, their sizes, and the moderators that change the effect sizes. Having measured effect sizes and identified important subpopulations “associated” with large or small effects, the next step should be to carefully measure the variance attributable to genuine individual differences in order to carefully define and evaluate the theory’s scope. In a comprehensive research strategy, these different components form a virtuous cycle. Having good effect size estimates thanks to meta-analyses and replication helps establish theoretical scope. Evaluating theoretical scope based on past studies also helps inform the design of future studies: By determining if a theory “applies to everyone” a priori, a replication will naturally become more efficient and generate more informative insights. Likewise, all of these components contribute to developing more concise and more accurate theories.

We advocate that future work dig more deeply into how different predictions are interrelated and how evidence can be combined across studies. For example, hierarchical models might sometimes spell out the joint distributions of multiple predictions, in which case the scholar may be able to move beyond the upper and lower bounds we have discussed. Another possibility is that some theories spell out theoretical mechanisms, cognitive channels, or contexts, according to which some experimental effects would show up for some individuals and in some experimental conditions, but not others (see, e.g., Jachimowicz, Duncan, Weber, & Johnson, 2019, in the example of “default effects” in decision making). To the extent that these theories predict conjunctions of disjunctions, such as

$$[\text{A OR B OR C}] \text{ AND } [\text{D OR E OR F}],$$

the paradox still applies to the conjunctions unless, in the sense that we have discussed before, these disjunctions can be rewritten as probability distributions over possible behaviors that apply to everyone.

In all, scholars should pay attention to important nuances in scientific discourse: Rather than claiming, vaguely, that certain types of treatments enhance "people's" memory, or reduce "people's" anxiety, and so forth, and that certain moderators enhance or reduce this "association," they should explicitly state who the theory applies to and whether the predicted "feature" in question is a probabilistic outcome.

References

Armor, D., Massey, C., & Sackett, A. (2008). Prescribed optimism: Is it right to be wrong about the future? *Psychological Science, 19,* 329–331.

Cohen, J. (1988). *Statistical power analyses for the social sciences*. Hillsdale, NJ: Erlbaum.

Ericsson, K. A., & Kintsch, W. (1995). Long-term working memory. *Psychological Review, 102,* 211–245.

Foroughi, C., Werner, N., Barragán, D., & Boehm-Davis, D. (2015). Interruptions disrupt reading comprehension. *Journal of Experimental Psychology: General, 144,* 704–709.

Foroughi, C., Werner, N., Barragán, D., & Boehm-Davis, D. (2016). "Interruptions disrupt reading comprehension": Correction to Foroughi et al. (2015). *Journal of Experimental Psychology: General, 145,* 881. http://dx.doi.org/10.1037/xge0000186

Guan, M., & Vandekerckhove, J. (2016). A Bayesian approach to mitigation of publication bias. *Psychonomic Bulletin & Review, 23,* 74–86.

Haaf, J. M., & Rouder, J. N. (2017). Developing constraint in Bayesian mixed models. *Psychological Methods, 22,* 779–798.

Jachimowicz, J. M., Duncan, S., Weber, E. U., & Johnson, E. J. (2019). When and why defaults influence decisions: A meta-analysis of default effects. *Behavioural Public Policy*. Advance online publication. http://dx.doi.org/10.1017/bpp.2018.43

Leavitt, J., & Christenfeld, N. (2011). Story spoilers don't spoil stories. *Psychological Science, 22,* 1152–1154.

McShane, B. B., & Böckenholt, U. (2017). Single-paper meta-analysis: Benefits for study summary, theory testing, and replicability. *Journal of Consumer Research, 43,* 1048–1063.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349,* aac4716.

Regenwetter, M., & Robinson, M. M. (2017). The construct-behavior gap in behavioral decision research: A challenge beyond replicability. *Psychological Review, 124,* 533–550.

Regenwetter, M., & Robinson, M. M. (2019). The *construct-behavior* gap revisited: Reply to Hertwig and Pleskac (2018). *Psychological Review, 126,* 451–454. http://dx.doi.org/10.1037/rev0000145

Ruscio, J. (2008). A probability-based measure of effect size: Robustness to base rates and other factors. *Psychological Methods, 13,* 19–30.

Schweinsberg, M., Madan, N., Vianello, M., & many others. (2016). The pipeline project: Pre-publication independent replications of a single laboratory's research pipeline. *Journal of Experimental Social Psychology, 66,* 55–67.

Shields, G. S., Sazma, M. A., McCullough, A. M., & Yonelinas, A. P. (2017). The effects of acute stress on episodic memory: A meta-analysis and integrative review. *Psychological Bulletin, 143,* 636–675.

Smith, P. L., & Little, D. R. (2018). Small is beautiful: In defense of the small-N design. *Psychonomic Bulletin & Review, 25,* 2083–2101.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science, 211,* 453–458.

Uhlmann, E., Pizarro, D., & Diermeier, D. (2015). A person-centered approach to moral judgment. *Perspectives on Psychological Science, 10,* 72–81.

van Assen, M. A., van Aert, R., & Wicherts, J. M. (2015). Meta-analysis using effect size distributions of only statistically significant studies. *Psychological Methods, 20,* 293–309.

van 't Veer, A., Lassetter, B., Brandt, M., & Mehta, P. (n.d.). Replication of prescribed optimism: Is it right to be wrong about the future? by David A. Armor, Cade Massey & Aaron M. Sackett (2008, Psychological Science). Retrieved from Osf.io/qlzap

Webb, N., Shavelson, R., & Haertel, E. (2006). Reliability coefficients and generalizability theory. In C. Rao & S. Sinharay (Eds.), *Handbook of statistics* (Vol. 26, pp. 81–124). the Netherlands: Elsevier.

Appendix

Definitions and Derivations

Throughout this Appendix, we assume that each participant has a true value of the hypothetical construct for both the treatment and control conditions. In the comparison of responses, we also assume that each respondent provides stochastically independent responses for both the treatment and the control condition. The purpose of these assumptions is to obtain easy to interpret and mathematically simple closed-form formulae. The various assumptions in this Appendix are purely for purposes of tractability and to be able to provide concrete numbers in our illustrations.

Comparing Responses

First, we consider differences of responses without attributing response variability to any particular source. Let $\mathbf{X}_T \sim N(\mu, \sigma^2)$ denote the responses in the treatment and $\mathbf{X}_C \sim N(\mu', \sigma^2)$ those in the control condition. For mathematical simplicity, we assume homogeneity of variance, and we assume that the two random variables $\mathbf{X}_T$ and $\mathbf{X}_C$ are independent. We consider the prediction $\mu > \mu'$ and the effect size measured by Cohen's d (Cohen, 1988)

$$d = \frac{\mu - \mu'}{\sigma}.$$

We can compare responses by considering the difference between $\mathbf{X}_T$ and $\mathbf{X}_C$. A draw from this variable, $\mathbf{\Delta_X} = (\mathbf{X}_T - \mathbf{X}_C) \sim N(\mu - \mu', 2\sigma^2)$, is positive if the participant's observed response in the treatment condition is larger than his or her observed response in the control condition. To calculate the probability that a randomly selected individual generates a larger response in the treatment condition than in the control condition, we need only calculate the probability of selecting an individual with $\mathbf{\Delta_X} > 0$. This probability is given by

$$Prob(\mathbf{\Delta_X} > 0) = \frac{1}{2\sqrt{\pi\sigma^2}} \int_0^\infty e^{-\frac{(x-(\mu-\mu'))^2}{4\sigma^2}} dx. \quad (4)$$

Multiplying $\mathbf{\Delta_X}$ by $\frac{1}{\sqrt{2}\sigma}$ and subtracting $\mu - \mu'$ yields a standard normal variable. Writing Φ for the cumulative density function for a standard normal distribution, Equation 4 then becomes

$$\begin{aligned} Prob(\mathbf{\Delta_X} > 0) &= Prob\left(\frac{\mathbf{\Delta_X} - (\mu - \mu')}{\sqrt{2}\sigma} > \frac{-(\mu - \mu')}{\sqrt{2}\sigma}\right) \\ &= \Phi\left(\frac{\mu - \mu'}{\sqrt{2}\sigma}\right) \\ &= \Phi\left(\frac{d}{\sqrt{2}}\right). \end{aligned}$$

If, instead, we considered a between subjects experimental design, under similar assumptions, then this calculation would be what is known as the "common language effect size" measure (Ruscio, 2008).

Incorporating Individual Differences and Measurement Error

We now consider a more detailed model where we partition the variance of the response distributions into two components: variance attributable to individual differences, denoted σ^2_{diff}, and variance due to measurement error, σ^2_{err}. We define each response variable as a sum of two independent and normally distributed random variables:

$$\mathbf{X}_T = \mathbf{H}_T + \mathbf{E},$$

$$\mathbf{X}_C = \mathbf{H}_C + \mathbf{E},$$

where $\mathbf{H}_T$ and $\mathbf{H}_C$ denote the value of the hypothetical construct in the treatment condition and control condition, and where the random variable $\mathbf{E}$ corresponding to measurement error has a mean of 0. Formally,

$$\mathbf{H}_T \sim N(\mu, \sigma^2_{\text{diff}}), \quad \mathbf{H}_C \sim N(\mu', \sigma^2_{\text{diff}}), \quad \mathbf{E} \sim N(0, \sigma^2_{\text{err}}).$$

Therefore, σ^2 becomes

$$\sigma^2 = \sigma^2_{\text{diff}} + \sigma^2_{\text{err}},$$

Δ_x becomes,

$$\mathbf{\Delta_X} = (\mathbf{X}_T - \mathbf{X}_C) \sim N(\mu - \mu', 2(\sigma^2_{\text{diff}} + \sigma^2_{\text{err}})),$$

and Cohen's d becomes

$$d = \frac{\mu - \mu'}{\sqrt{\sigma^2_{\text{diff}} + \sigma^2_{\text{err}}}}.$$

We are interested in a person's differences in value of the hypothetical constructs between the treatment and control conditions, which we denote $\mathbf{\Delta_H}$, and which is distributed as

$$\mathbf{\Delta_H} = (\mathbf{H}_T - \mathbf{H}_C) \sim N(\mu - \mu', 2\sigma^2_{\text{diff}}).$$

If there are no individual differences, i.e., if $\sigma^2_{\text{diff}} = 0$, then $\mathbf{\Delta_H}$ is the constant $\mu - \mu'$. If there is no measurement error, that is, if $\sigma^2_{\text{err}} = 0$, then $\mathbf{\Delta_H} = \mathbf{\Delta_X}$. Let q denote the proportion of total variance attributable to individual differences in the population, that is,

$$q = \frac{\sigma^2_{\text{diff}}}{\sigma^2_{\text{diff}} + \sigma^2_{\text{err}}} = \frac{\sigma^2_{\text{diff}}}{\sigma^2}.$$

In all, we obtain the probability $Prob(\mathbf{\Delta_H} > 0)$ that a randomly selected individual has a higher value on the hypothetical construct in the treatment condition than in the control by standardizing $\mathbf{\Delta_H}$:

$$\begin{aligned} Prob(\mathbf{\Delta_H} > 0) &= Prob\left(\frac{\mathbf{\Delta_H} - (\mu - \mu')}{\sqrt{2}\sigma_{\text{diff}}} > \frac{-(\mu - \mu')}{\sqrt{2}\sigma_{\text{diff}}}\right) \\ &= \Phi\left(\frac{\mu - \mu'}{\sqrt{2}\sigma_{\text{diff}}}\right) \\ &= \Phi\left(\frac{\mu - \mu'}{\sigma} \times \frac{1}{\sqrt{2}} \times \frac{\sigma}{\sigma_{\text{diff}}}\right) \\ &= \Phi\left(\frac{d}{\sqrt{2q}}\right). \end{aligned}$$

Received January 9, 2019
Revision received April 26, 2019
Accepted May 5, 2019 ■