Social and Personality Psychology Compass 9/4 (2015): 188–201, 10.1111/spc3.12166

In Search Of True Things Worth Knowing: Considerations For A New Article Prototype

Alexa M. Tullett*
University of Alabama

Abstract

Within social and personality psychology, the existing "old prototype" of a publishable article is at odds with new expectations for transparent reporting. If researchers anticipate having to report everything while continuing to aim for a research product that includes multiple studies, examining a novel effect, with only statistically significant results, this will have negative implications for initial decisions about what research to conduct. First, researchers will be discouraged from collecting additional data because this could potentially mar existing findings. Second, they will be discouraged from pursuing questions for which the answers are unknown, as this would be a waste if the results do not fit old-prototype expectations. These practices undermine what seem to be two universal values within personality and social psychology: truth and interestingness. Suggestions for a "new prototype" that de-emphasizes *p*-value cutoffs, multiple studies, and novelty will be discussed with an eye toward encouraging research decisions that foster true and interesting findings.

Social and personality psychologists exhibit some reluctance about transparently reporting their scientific findings (Fuchs, Jenny, & Fiedler, 2012; John, Loewenstein, & Prelec, 2012). Although it might be tempting to attribute this hesitation to laziness or lack of integrity, it could be informative to understand the source of this concern. Even if psychologists recognize that transparent reporting provides an important safeguard against false positives, psychologists likely also recognize that our field prizes a particular type of research product that I will optimistically refer to as the "old prototype". That is, good papers are expected to describe a (sometimes large) set of studies that all show statistically significant variations of a novel effect and no nonsignificant variations of that effect. One possible reason for researchers' reluctance to report all of the messy details of their work, then, is that expectations for transparency and expectations for old-prototype-style papers are at odds with each other; the old prototype is largely inconsistent with what transparently reported science looks like (Miguel et al., 2014).

This conflict has the potential to have a negative effect not just on the way people analyze and report data but also on the way people go about collecting data in the first place. If a researcher anticipates writing an old-prototype-style paper while reporting all relevant information – from independent variables and dependent variables to pilot studies and replications – then this will naturally impact decisions about what data to pursue. Specifically, researchers will be deterred from gathering additional data, in the form of an extra measure, more participants, or an additional study, in case it "messes up" an existing set of findings. In addition, researchers will be discouraged from pursuing highly original ideas for fear that reporting preliminary null results will undermine any more robust findings obtained down the line. In other words, trying to balance transparency and old-prototype expectations can lead to decisions that are clearly unscientific. If one accepts the old prototype, or if one thinks it is unlikely to change, then it would seem understandable to be concerned about

requirements for transparency. On the other hand, if the old prototype evolves, then perhaps researchers will be more willing, and even eager, to move toward greater transparency and thus reduced bias in the psychological literature.

Other authors have suggested that a revised prototype is needed. Simmons, Nelson, and Simonsohn (2011) suggested that journals must tolerate messier data if they are to accommodate greater transparency. This sentiment has been echoed by Giner-Sorolla (2012) who notes that the prettiness of a narrative should not be the primary basis for evaluation of scientific papers, by Kaiser (2012) who proposes a "campaign for real data", and by King (2012) who emphasizes the importance of a "true story" over a "good story". Extending these ideas, Maner (2014) has recommended concrete ways in which reviewers and editors can encourage "real data" without lowering scientific standards.

Here, I will attempt to contextualize these suggestions within a broader discussion of the core values of social and personality psychology. I will begin by taking a step back to discuss what I consider to be two important and largely universal goals within the field: truth and interestingness. Then I will attempt to document how transparency is necessary for these goals and how the old prototype can undermine these goals, even if transparent reporting is universally adopted, by influencing the research that people decide to conduct. Finally, I will provide suggestions for a revised prototype with the hopes that a new mental representation will encourage practices that make social and personality research more true and more interesting.

Universal Values

In order to evaluate what a new prototype should look like, there must be some consensus about which research practices should be incentivized and, at a more fundamental level, which basic values should guide psychological research. Although disagreements about the particulars of best research practices are not uncommon, there seem to be some core values that are largely uncontroversial (Funder et al., 2014). Specifically, regardless of the methodological, theoretical, or statistical camp, both truth and interestingness emerge as important values within social and personality psychology.

Truth

Uncovering truths about the world is arguably the primary goal of science (Boyd, 1983; Smart, 1963; van Fraassen, 1980). This can be contrasted with two other outcomes: uncovering untruths ("false positives" or Type I error) and failing to uncover truths ("misses" or Type II error). Importantly, uncovering truths encompasses both correctly identifying effects that exist ("hits") and correctly rejecting effects that do not exist ("correct rejections"). For this reason, claiming that truth should be valued in the scientific publication process is not the same as saying that findings should only be published when we are almost certain that they are nonzero (as would be implied by a very low *p*-value). Instead, it amounts to saying that research should be rewarded to the extent that it enhances our understanding of an effect, regardless of whether that effect is close to or far from zero (Cohen, 1994; Cumming & Finch, 2005; Greenwald, 1975; Nickerson, 2000).

Reasons for valuing truth extend beyond the obvious. Perhaps most importantly, having an accurate and precise understanding of effects is necessary for using psychological findings as the basis for future predictions, which is one of the hallmarks of scientific investigation (Colyvan, 2001; Popper, 1963). Furthermore, the confidence that we can have in an initial finding necessarily sets a limit on the confidence that we can have in findings that depend on it. For instance, a study that uses self-reports of behavior as a proxy for actual behavior is only as informative as the existing knowledge of the association between the two. Of course, the study

of human behavior is complicated, and useful contributions can be made without the kind of exactness that might be expected in chemistry or physics. Nevertheless, too high a degree of imprecision precludes the ability to develop effective interventions, evaluate the potential consequences of real-world events, or conduct cumulative lines of work.

Interestingness

Valuing truth seems relatively uncontroversial, but if it were the only thing prioritized by psychologists, there could be consequences that even the staunchest truth advocate might regret. If all we cared about were small confidence intervals, psychologists would only pursue effects that are extremely easy to assess (e.g., very large effects or effects that can be assessed with quick and inexpensive methodology). Clearly, we also care, at least to some degree, about whether or not the effect in question is something worth knowing.

It seems apparent, then, that psychological researchers value interestingness (Gray & Wegner, 2013). When I refer to a finding as interesting, I mean to indicate that it provides knowledge that has some kind of value. This value could come from its practical utility, its ability to satisfy basic curiosity, its incremental contribution to a broader body of work, or potentially a range of other sources. Although this definition is certainly subjective, it allows for the label "interesting" to be applied to both questions that are applied and basic, both results that are novel and non-novel, and both findings that are fascinating and mind numbing to one's grandmother. It also implies that a finding can fall below a standard of interestingness, regardless of how true it may be, if it does not provide information that is sufficiently valuable.

Claiming that researchers value interestingness is almost tautological; people are more likely to care about things that are interesting than things that are uninteresting, by definition. This is slightly different, however, than saying that we should give more interesting findings higher priority than less interesting findings. Despite compelling reasons to think that interestingness is currently overvalued (see section on Novelty), it seems obvious that this criterion should not be ignored completely. Of course, what counts as "interesting" will likely remain a source of contention (Bornmann, Mutz, & Daniel, 2010; Petty, Fleming, & Fabrigar, 1999), but if psychologists generally agree that interestingness is valuable, then there are implications for how best to incentivize research efforts.

One caveat to the preceding discussion of values is that, even if truth and interestingness are both important, interestingness depends on truth in a way that is not reciprocal. For example, consider the hypothetical finding that cheating on your spouse increases marital satisfaction. This might be fairly interesting, but it becomes far less interesting if it turns out to be untrue. Conversely, consider the hypothetical finding that people feel more frustrated when they get an F on a test than when they get an A. Even if this is completely uninteresting, this does not in any way make it less true. This asymmetry suggests that, although it may be a mistake to disregard interestingness, disregarding truth is more fundamentally problematic.

If there is some consensus that truth and interestingness (but particularly truth) are important values within social and personality psychology, then it would seem important to align expectations for research output with these values. Currently, old-prototype expectations can discourage research practices that are necessary for upholding these values. Specifically, they can discourage researchers from collecting data that could challenge existing findings and from testing hypotheses that might be wrong. One potential solution, then, is to restructure expectations for publishable manuscripts in a way that incentivizes research practices likely to lead to true and interesting findings.

Expectations Created by the Old Prototype

Uncontroversial expectations

Many of the things valued by the current publication system are things that we should continue to value. For instance, reviewers and editors are likely to dismiss papers unless measures are valid and reliable, experimental manipulations are demonstrably effective, and interpretations follow closely from the data while considering, if not entirely ruling out, alternative explanations. I consider these expectations to be uncontroversial requirements for good scientific publications.

Another uncontroversial expectation is high statistical power. Increasing statistical power decreases the risk of Type II error, decreases the proportion of published results that are Type I error, and improves the precision of effect size estimates (Cohen, 1988, 1992; Fraley & Vazire, 2014). In other words, statistical power is indispensable to the goal of truth. Although there is often a trade-off between the resource intensiveness of a method and the statistical power that can realistically be achieved (i.e., some compromise in power is necessary if one uses costly or time-consuming methods), putting a lower limit on power is critical to ensuring confidence in published effects (Bakker, van Dijk, & Wicherts, 2012; Maxwell, 2004). This point has led some to suggest that if there is a problem with current expectations for statistical power, it is that they are not demanding enough (Fraley & Vazire, 2014; Schimmack, 2012; Sedlmeier & Gigerenzer, 1989).

Transparency. Recently, some journals have begun increasing their expectations for transparent reporting. These requirements have been introduced to combat the motivation (encouraged by the old prototype) to selectively report results in order to tell the cleanest story. For instance, *Psychological Science* and *Social Cognition* have introduced disclosure statements that require authors to list all measures, all manipulations (if any), all excluded participants, and a priori rationales for sample size. More stringent expectations could involve reporting results of pilot studies as well as any replications (conceptual or direct) that have been conducted by the authors. They could also involve making data, analyses, and hypotheses publically available (Nosek & Bar-Anan, 2012). Although there are ongoing debates about the practicalities of implementing some of these expectations (e.g., Asendorpf, 2012 and Spellman, 2012), it seems that greater transparency can only lead to greater understanding and that recognizing this has important implications for how a new prototype should look.

It is hard to overestimate the importance of transparent reporting to the goal of pursuing truth (Miguel et al., 2014). Consider the following thought experiment. You are given an opportunity to bet money on the outcome of a replication study, and you have two options: (i) you can either read the original old-prototype-style manuscript, or (ii) you can see all of the relevant data files. Of course, you would have a much better chance of making money if you chose Option 2. Choosing Option 1 would be a bad way to make predictions because Option 1 compromises on transparency, which is necessary for truth, which is in turn necessary for accurate prediction.

If transparency is required for truth and truth is a basic value in social and personality psychology, then this suggests that transparency should be a high priority. Furthermore, transparency is also required for interestingness to the degree that interestingness depends on truth (N.B. Transparency could take away from the interestingness of an article's *narrative*, but this is distinct from detracting from the interestingness of a *finding*). Thus, if truth and interestingness are important values, transparency is a necessary expectation.

Consistent with this sentiment, others have explored the consequences that have arisen as a result of nontransparent, or selective, reporting. These analyses have focused on estimating the rate of false positives (one type of untruth) in the literature (Ioannidis, 2005) and on demonstrating how selective reporting contributes to this problem (John et al., 2012; Simmons

et al., 2011). What is particularly worrisome is that a lack of transparency not only leads to false positives, it also precludes the identification of these false positives. The most serious concerns about the current state of social and personality psychology have arguably arisen not because there may be false positives in the literature but because we are unable to identify which findings fall into this category (Nosek, Spies, & Motyl, 2012; Wagenmakers, Wetzels, Borsboom, van der Maas, & Kievit, 2012).

One criticism of requirements for greater transparency stems from the concern that complete transparency is impossible because we can never know all of the details of a study. Even if all materials, procedures, and analyses were made available, there would still be aspects of the study – such as the temperature in the lab room or the appearance of the experimenters – that would remain unknown to the reader. Thus, researchers will inevitably be forced to make subjective judgments about which details must be included and which can be safely ignored. In tackling this problem, other authors have posed carefully considered, concrete recommendations for how to achieve the types of transparency that are most critical (Nosek & Bar-Anan, 2012; Simmons et al., 2011). Although pragmatics may prevent complete transparency from being attainable, practical limitations do not seem so overwhelming that they negate the usefulness of transparency as an ideal, just as the impossibility of identifying a population effect size does not negate the usefulness of precision as an ideal.

Problematic expectations

A number of the expectations of the current publication system can come into conflict with the values of truth and interestingness. Some of these have served as imperfect proxies for these values (e.g., a *p*-value less than .05 means "true" or a novel finding must be "interesting"). Here, I will discuss ways in which these expectations can conflict with these values and thereby make a case for how loosening these old-prototype expectations could reduce some of their adverse motivational consequences.

p-values under .05. Surveying the range of *p*-values present in published articles quickly makes it clear that within the current publication system, there is an expectation for *p*-values under .05 (Fanelli 2010, 2012; Greenwald, 1975; Rosenthal, 1979; Sterling, 1959; Sterling, Rosenbaum, & Weinkam, 1995). This expectation arose as a way to enforce truth. When reported as an exact value (as opposed to the categorical statement $p < .05$) alongside an effect size, *p*-values tell us about the range of population effect sizes that are plausible given the observed results (Nickerson, 2000). In principle, prioritizing small *p*-values reduces Type I error and increases the precision with which we estimate effects.

p-values can become problematic, however, when thresholds such as $p < .05$ are identified and enforced within the publishing framework. Because it is more difficult to publish findings that have *p*-values exceeding .05 (Fanelli, 2012; Greenwald, 1975), researchers are motivated to find statistically significant results and are consequently more likely to engage in practices that artificially decrease their *p*-values (Simmons et al., 2011). This has led some to estimate that a large proportion of findings reported in the scientific literature may not be replicable (Ferguson & Heene, 2012; Ioannidis, 2005; Ioannidis, 2008).

More stringent expectations for transparency address many of the problems that arise because of the prioritization of statistical significance. These expectations limit selective reporting and thus help to minimize "*p*-hacking" – the process of capitalizing on chance created by multiple analysis options (Simmons et al., 2011). The deeper problem with placing such a strong emphasis on arbitrary cutoffs for significance, however, is that even if researchers are honest about their data, they will still *want p*-values less than .05. This creates a conflict of interest as the motivation to publish is at odds with being agnostically open to any result (Nosek et al., 2012). In

conjunction with expectations for transparency, this preoccupation with statistical significance can discourage people from collecting more data when they already have significant results. It is a risky move to conduct a replication (particularly a direct replication) when a nonsignificant result will markedly lower the publishability of a paper. Paradoxically, then, this emphasis on the .05 cutoff can lead to practices – and more broadly a mental set – that obscures truth.

Requiring statistical significance can also undermine interestingness. It is not difficult to obtain *p*-values under .05, it is just difficult to do so for things that people would be interested in knowing. For example, the correlation between responses to the statements "I appreciate music" and "I like to listen to music" would likely be highly significant even with relatively few participants. If publishable papers can contain only significant results, and if people are expected to report all of their results, then the logical approach to take as a researcher is to pursue very "safe", obvious effects (although this would likely fail to satisfy current expectations for novelty discussed below). If, on the other hand, it were permissible to publish a paper that had nonsignificant results, then people could reasonably devote time and resources to looking for an effect that may or may not be different than zero. Presumably, if worthwhile science starts from hypotheses that could feasibly be wrong, this practice should characterize a substantial proportion of scientific work (Popper, 1963).

In sum, if more stringent requirements for transparency are adopted, as is arguably required if truth is an important value, then having a *p*-value cutoff creates some problematic incentives. If existing data are statistically significant, then researchers are discouraged from collecting more data, which is inconsistent with valuing truth. If data have not yet been collected, then researchers are discouraged from pursuing questions to which they don't already know the answer, which is inconsistent with valuing interestingness.

Multiple studies. Most journals reward, if not require, multiple studies. Like *p*-value cutoffs, this expectation arose as a way to enforce truth (Schimmack, 2012). If any individual study could be an instance of Type I error (i.e., a "fluke"), then it would seem reasonable to require researchers to find an effect multiple times before that finding ends up in a journal (Cesario, 2014). After all, if an effect is truly zero, then the chance of erroneously concluding it is different than zero in one study is much higher (5%) than the chance of making this incorrect conclusion in three consecutive studies (.0125%).

All else being equal, having more studies does indeed offer multiple benefits, such as providing evidence of replicability, identifying moderators, and establishing generalizability. In practice, however, it might not be the case that the expectation of multiple studies results in researchers doubling or tripling the amount of data they would have collected had one study been enough (Cohen, 1992; Maxwell, 2004; Rossi, 1990; Schimmack, 2012; Sedlmeier & Gigerenzer, 1989). Assuming resources are limited (e.g., 500 participants can feasibly be recruited for a project), the researcher is faced with a decision about how to allocate those resources. If, in order to fulfill the multiple studies expectation, the researcher decides to conduct three studies with these participants instead of one, this has the consequence of undercutting statistical power in each and increasing the number of statistical tests, thereby increasing the rate of Type I error (Bakker et al., 2012; Cohen, 1962; Maxwell, 2004; Schimmack, 2012). Although there are times it might make sense to run three studies instead of one (e.g., if the predicted effect size is relatively large and the compromise in statistical power is offset by the ability to test boundary conditions), the rule "two studies is better than one" could be counterproductive if applied indiscriminately.

Another consequence of the expectation for multiple studies is that it magnifies the problems associated with *p*-value cutoffs. The old prototype entails an expectation of "tidiness"; any predicted effect, whether it be a main effect, interaction, or simple effect, should have a *p*-value less

than .05, whereas any unpredicted effect should have a *p*-value above .05 (or, preferably, above the "marginal" .1 value; Maner, 2014). This is not likely even when studies are reasonably well powered. For instance, even with 80% power per study, there is only a 33% chance of finding significant results in five consecutive studies (see Schimmack, 2012, Figure 1). It is even less likely if we anticipate that people will sometimes overestimate effect sizes, underestimate sources of error, or be wrong about moderators (Ioannidis, 2008; Young, Ioannidis, & Al-Ubaydli, 2008). If a researcher is just beginning a line of research, it would be risky to pursue an effect that might not exist (i.e., an interesting effect) if one or two nonsignificant results will make the entire endeavor unpublishable.

An important caveat to the preceding section is that testing a given effect multiple times is an extremely important practice as a field (Frank & Saxe, 2012; Koole & Lakens, 2012; Nosek et al., 2012; Simons, 2014). Replication is necessary for building confidence in effects and understanding the limits of those effects (see Novelty section below). But, the expectation that manuscripts are *always* better if they contain multiple studies can conflict with the values of truth and interestingness. It can conflict with truth by prompting researchers to divide their resources in ways that increase Type I error. It can conflict with interestingness by compounding problems caused by the expectation of statistical significance; if all results are to be reported and all have to be significant, then nonobvious effects become very risky investments of time and resources.

Novelty. It may seem inconsistent to claim that interestingness is a universal value and to then question the importance of novelty. Previously, I have suggested that researchers need to have the freedom to pursue new ideas if they are to make interesting discoveries. Novelty and interestingness, however, are not synonymous. For some of the field's most prestigious journals, the term "novel" is reserved for tests of ideas that have never been explored before and not for research that allows for more precise estimation of effects (e.g., direct replications) or research that establishes boundary conditions (e.g., conceptual replications; Srivastava, 2012). Arguably, this high bar may be justified at the most widely read publication outlets, but even more lenient standards (e.g., refusing to accept studies that test an existing idea without theoretical extension) deter research aimed at establishing the robustness of influential ideas. It seems beneficial to prioritize studies that tell us something new, but if every paper must sufficiently distinguish its theoretical contribution from those that came before, then this has the potential to discourage important work.

The problem with a universal requirement for novelty is that it conflicts with the goal of truth (Srivastava, 2012). Any individual study – even ones that are high powered and rigorous – has the potential to be a false positive (Ledgerwood & Sherman, 2012). Testing an idea multiple times (even if not in a single manuscript) is crucial to understanding an effect, whether it be by retesting exactly the same idea using a direct replication or by testing the boundaries of an idea using a conceptual replication.

Earlier, I suggested that truth, as a singular value, is not enough; a study should also be interesting if it is to constitute a meaningful contribution to the field. But, compromising novelty does not necessarily mean compromising interestingness. One reason for this is that it may be relatively common to overestimate the obviousness of effects (Button et al., 2013; Nosek et al., 2012). This is even more likely if, as some suggest, the current rate of false positives in the literature is quite high (Ioannidis, 2005, 2008). If we take seriously the idea that robust effects are harder to come by than we originally thought, then even seemingly obvious effects become more impressive (Most people would likely be surprised, for instance, to learn that you would need approximately 47 people per cell to detect an association between liking eggs and eating egg salad; Simmons, 2014). Given that establishing an effect is likely a harder task than we have

come to believe, even direct replications are interesting in so far as they provide us with more confidence in an interesting finding.

A New Prototype?

If truth and interestingness are indeed core values of social and personality psychology, an important implication of this is that we cannot continue to expect manuscripts to fit the old prototype. This does not entail *lowering* expectations for manuscripts; instead, it involves changing expectations in ways that allow researchers to prioritize these two values while not compromising the publishability of their findings (Maner, 2014). One important step in this direction, and one that is already underway, is increasing expectations for transparent reporting. It is impossible to prioritize truth, arguably the most fundamental value, without putting limits on selective reporting.

In light of these new expectations for transparency, rigidly upholding the old prototype has the potential to cause researchers to make significant compromises when it comes to truth and interestingness. If researchers are to embrace transparent reporting but are also expected to produce manuscripts where they show a novel effect, across multiple studies, with all findings reaching traditional levels of statistical significance, then this will have a decidedly negative impact on the research they decide to conduct. Broadly speaking, they will likely be less motivated to collect extra data, be it extra measures, extra participants, or extra studies, because this could potentially "mess up" existing findings. They will also likely be less motivated to pursue effects when they are not confident that they will find them. Because these motivations are in conflict with truth and interestingness, it seems that a new prototype – one that does not place the same emphasis on nominal significance, multiple studies, and novelty – is necessary.

De-emphasizing p-values

One change that could potentially have a dramatic effect on the way that people conduct research would be to loosen the requirement that *p*-values need to be lower than .05. This would lead to at least two further changes. First, it would be easier to publish null findings that call into question a previously reported effect size (e.g., a "failed" replication). Second, and perhaps more controversially, it would be easier to publish interesting but inconclusive research, for instance, a new line of studies in which some results are significant and others are not. At a more fundamental level, if researchers were able to publish findings that were not clearly different than zero, this would allow them to publish answers to interesting questions regardless of what those answers might be (Greenwald, 1975; Fanelli, 2012).

Allowing definitive $p > .05$. Sometimes, perhaps frequently, "no" is an interesting and important answer to a scientific question. For instance, does a particular intervention cause reductions in violent crime? Allowing *p*-values greater than .05 would allow for the publication of persuasive null results. This is important if there is to be any avenue for identifying false positives in the literature (Greenwald, 1975; Rosenthal, 1979). It is also important for ameliorating the file-drawer problem, thus providing more accurate meta-analytic estimates of effects (Ioannidis, 2008; Rosenthal, 1979). Finally, if publishing these results is possible, then this provides an outlet for disseminating null answers to interesting and previously unexplored questions.

One potential objection to the publication of null results is the argument that these findings are difficult to interpret (Mitchell, 2014; Stroebe & Strack, 2014). This is a complex problem and one that goes beyond the scope of the current paper, but there are a number of authors who have proposed promising solutions including increasing statistical power, using manipulation checks, or departing from null hypothesis significance testing in favor of Bayesian techniques (Cohen, 1962; Edwards, Lindman, & Savage, 1963; Greenwald, 1975; Rossi, 1990).

Exploring strategies for improving the interpretability of null effects could enhance the self-correctional potential of psychology and thus improve confidence in published findings (Stroebe, Postmes, & Spears, 2012).

Allowing ambiguous $p > .05$. Allowing *p*-values greater than .05 would also allow for the publication of ambiguous and perhaps suggestive results. As discussed above, many of those in favor of publishing null results are willing to allow *p*-values greater than .05 when these are interpreted as evidence that an effect *does not* exist. In other cases, however, *p*-values greater than .05 will be interpreted as weak evidence that an effect *does* exist. If people become more cautious in drawing definitive conclusions from individual studies (Ledgerwood, 2014; Braver, Thoemmes, & Rosenthal, 2014), then these findings have the potential to make a valuable contribution (e.g., exposing other researchers to an idea or providing preliminary evidence about effective operationalization). If studies in this category were more publishable, this might encourage researchers to include studies that were inconclusive alongside studies that were relatively compelling, providing readers with valuable information about the robustness of the effect and about the consequences of different methodological choices. Moreover, it would provide incentive for pursuing original ideas and disseminating groundbreaking (but not definitive) results (Srivastava, 2012).

An important objection to publishing *p*-values that hover around the .05 mark stems from concerns about the informational value of these results. If our main concern is whether or not an effect is different from zero, these marginally significant results – those where the confidence interval neither clearly includes nor excludes zero – are relatively unsatisfying. Furthermore, these effects currently seem to be the least replicable, based on both empirical and theoretical analyses (Gilbert, 2014; Lakens & Evers, 2014). In some instances, it seems that results are so ambiguous that they leave readers in some position they were in before reading the study: unsure of whether or not the effect exists.

The challenge that arises, then, is that of differentiating between a study that is informative but inconclusive and one that is too imprecise to provide any new information. Here, I think it could be helpful to consider whether the answer to the research question is only interesting if it is a concrete "yes" or "no". If this is the case, then it may make sense to evaluate these manuscripts on the definitiveness of the answer provided (an evaluation that would involve considerations of statistical power, methodological rigor, etc.). If this is not the case, however, then evaluation of the methods independent of the results (a possibility that I discuss further below) seems the most sensible way to evaluate the study. This scenario may be more common that it initially seems. It may characterize studies that are included within larger papers that, on the whole, provide compelling evidence of an effect, studies that resolve confounds in previously published studies that report robust effects, studies that investigate an exciting idea for the first time and thus primarily contribute an interesting question, or studies that investigate questions where the precision of the effect size estimate is of greater concern than whether or not it is different from zero. For these reasons, allowing publication of ambiguous findings where $p > .05$ has the potential to provide new information, at least in some cases.

De-emphasizing multiple studies

As recommended previously by others (e.g., Bakker et al., 2012 and Schimmack, 2012), loosening the expectation of multiple studies could have a beneficial influence on research practices. Importantly, this is not the same as proposing that researchers conduct fewer replication studies. Instead, it amounts to suggesting that researchers not be required to divide resources across a series of studies in cases where it would make more sense to conduct

one more resource intensive study. Such a change would promote truth by incentivizing high-powered studies, which offer greater confidence in effects. It would also promote interestingness by making it more worthwhile to invest in data that are difficult to collect, for instance, longitudinal, psychophysiological, neuroscientific, or behavioral data (Baumeister, Vohs, & Funder, 2007).

A possible concern with allowing more single-study papers is that one-shot false positives will be more likely to be published. Indeed, if expectations for statistical significance and novelty remain the same, such a change could be detrimental. If single studies can be published but replications cannot, then loosening the expectation for multiple studies could inflate the Type I error rate and leave no avenue for correction (Ferguson & Heene, 2012; Ioannidis, 2012; Stroebe et al., 2012). For this reason, the efficacy of this change is partially contingent on adjusting expectations regarding *p*-values and novelty. Optimistically speaking, though, concurrent changes in these expectations would allow a segment of researchers to devote resources to individual, high-powered, and/or methodologically complex studies, which have the potential to contribute to the field in ways that smaller, simpler studies cannot.

It is worth considering that the old-prototype expectation of multiple studies is only unrealistic if we accept a basic assumption: resources are limited (Schimmack, 2012; Srivastava, 2012). This might be an unwarranted assumption if we consider the possibility of either increasing efficiency or slowing the rate of publication (Nelson, Simmons, & Simonsohn, 2012). For instance, researchers might be able to accomplish this goal by running only very high-powered studies and only publishing definitive results (null or non-null). Indeed, publishing less would decrease the pressure to sacrifice quantity of participants for quantity of studies. As a result, the expectation of multiple studies could be maintained and along with it the benefits of conceptual replication.

De-emphasizing novelty

The benefit of loosening expectations of novelty is that it frees researchers to fruitfully invest in conducting replications (Koole & Lakens, 2012). Here, when I suggest that novelty should be de-emphasized, I do not mean that a study should not be required to provide new information; rather, I mean that a study should not be required to examine an idea that has not previously been explored. Replications are critical to the goal of truth – direct replications improve our understanding of the size of an effect, and conceptual replications add to our knowledge of the boundaries of an effect. Pursuing this knowledge is not in researchers' best interests unless non-novel studies are valued.

Another benefit of allowing, or even encouraging, more redundancy across studies is that it could facilitate greater theoretical integration. Spellman (2012) has noted that advancing as a predictive science requires integrating the increasingly massive amount of information we collect. Similarly, calls for "paradigm-driven research" (Nosek et al., 2012) and greater investment in theory development (Ledgerwood & Sherman, 2012; Monroe, 2014) highlight the crucial role that theory plays in building cumulative lines of work.

A possible objection to de-emphasizing novelty is that it could lower the bar for interestingness. This concern seems manageable for two reasons. First, if non-novel research (i.e., research on ideas that have been previously investigated) is still required to tell us something new, then this change does not actually abandon expectations for interestingness, particularly if we acknowledge that it is easy to overestimate the obviousness of effects. Second, if interestingness is dependent on truth (i.e., something is only interesting in so far as it is true), then it is necessary to prioritize truth, even above interestingness.

In sum: emphasizing methods rather than results

The above suggestions, at their root, call for a shift in focus from answers to questions (Vazire, 2014). Several authors have noted that one powerful way to incentivize good research practices is to evaluate research projects based on their methods rather than their results (Giner-Sorolla, 2012; Nosek et al., 2012). Another way to think of this is that researchers embarking on a new project, as well as reviewers and editors evaluating a completed project, should start out by asking themselves the following question: "How excited would I be to see these data?"

Critics of such an approach have posited that significant results are an important indicator of the quality of the methods. By analogy, most people would evaluate a cake by tasting it rather than by analyzing the recipe. Indeed, if the goal of a study is to establish the existence of an effect, then evaluating the results does seem the most efficient way to see if this goal has been achieved. If, on the other hand, the goal of a study is to answer a question about whether or not an effect exists or about the nature of an effect, then evaluating the methods (including results, like those of manipulation checks, that are designed to validate methods rather than answer a focal question) appears to provide the best way to decide whether the answer that is achieved is meaningful. The reason that a significant result is not necessarily an indication of sound methodology is that, although errors and chance can certainly make positive results null, errors and chance can also make null results positive (Neuroskeptic, 2014). If we are asking questions with unknown answers, it seems that good answers can only be distinguished from bad ones by knowing how they were obtained.

Conclusion

One potential reason that requirements for transparency can be threatening is that the old manuscript prototype is inconsistent with these new expectations. A consequence of this is that even people who value transparent reporting are justifiably worried that their honesty will make their findings unpublishable (Maner, 2014). Of course, when this honesty reveals fatal flaws in the research, an editor's decision to reject a manuscript can be a desirable outcome for the field (even if it is undesirable for the individual researcher). In many circumstances, however, findings will be dismissed not for fatal flaws but for imperfections (Simmons et al., 2011; Schimmack, 2012). On the contrary, transparently reported findings that include some degree of messiness will typify new manuscripts if we care about truth and interestingness.

An alternative way to respond to the conflict between old prototype expectations and transparency would be to change the way studies are designed from the outset, with the goal of conducting research that has nothing to hide. Although this sounds like an admirable and worthwhile goal, when one considers the logical fallout of such a strategy under the current norms, it starts to appear less promising. Pursuing the old prototype while following guidelines for transparency creates perverse incentives and, perhaps more troublingly, may not always be possible.

Here, I have suggested that truth and interestingness are universal values within social and personality psychology. If increased expectations for transparency are important for truth, then the old prototype needs to change to accommodate this transparency. In particular, putting less of an emphasis on *p*-value cutoffs, multiple studies, and novelty would allow researchers to conduct research without worrying about creating a superficially perfect research package. With these changes, perhaps researchers will be less constrained by the old prototype and have more freedom to devote their efforts to answering interesting questions.

Short Biography

Alexa Tullett's research focuses on how people make sense of the world. This interest is manifest in two main lines of research, one examining the beliefs that people use to organize the world and the other examining the ways in which people come to understand the thoughts and feelings of others. To explore these issues, her lab employs a social neuroscience approach, using neuroscientific, psychophysiological, self-report, and behavioral measures in an attempt to obtain a comprehensive understanding of social psychological phenomena. She has authored or co-authored papers in these areas for *Perspectives on Psychological Science*, *Social Cognitive and Affective Neuroscience*, and *Psychological Science*. Tullett holds a BSc in Physiology and Psychology from Western University and a PhD in Psychology from the University of Toronto. Currently, she teaches at the University of Alabama.

Note

* Correspondence: Psychology, University of Alabama, 505 Hackberry Lane, Tuscaloosa, AL 35478-0384, USA. Email: alexa.tullett@gmail.com

References

Asendorpf, J. B. (2012). Does open scientific communication increase the quality of knowledge? *Psychological Inquiry*, **23**, 248–250.

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science*, **7**, 543–554.

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

Bornmann, L., Mutz, R., & Daniel, H. D. (2010). A reliability-generalization study of journal peer reviews: A multilevel meta-analysis of inter-rater reliability and its determinants. *PLoS ONE*, **5**(12), e14331.

Boyd, R. N. (1983). On the current status of the issue of scientific realism. *Erkenntnis*, **19**, 45–90.

Braver, S. L., Thoemmes, F. J., & Rosenthal, R. (2014). Continuously cumulating meta-analysis and replicability. *Perspectives on Psychological Science*, **9**(3), 333–342.

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafo, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, **14**, 365–376.

Cesario, J. (2014). Priming, replication, and the hardest science. *Perspectives on Psychological Science*, **9**, 40–48.

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

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd edn). Hillsdale, NJ: Erlbaum.

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

Cohen, J. (1994). The Earth is round ($p < .05$). *American Psychologist*, **49**, 997–1003.

Colyvan, M. (2001). *The Indispensability of Mathematics*. New York, NY: Oxford University Press.

Cumming, G., & Finch, S. (2005). Inference by eye: Confidence intervals and how to read pictures of data. *American Psychologist*, **60**, 170–180.

Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, **70**, 193–242.

Fanelli, D. (2010). "Positive" results increase down the hierarchy of the sciences. *PLoS ONE*, **5**(4), e10068.

Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries. *Scientometrics*, **90**, 891–904.

Ferguson, C. J., & Heene, M. (2012). A vast graveyard of undead theories: Publication bias and psychological science's aversion to the null. *Perspectives on Psychological Science*, **7**, 555–561.

Fraley, R. C., & Vazire, S. (2014). The N-pact factor: Evaluating the quality of empirical journals with respect to sample size and statistical power. *Manuscript in preparation*.

Frank, M. C., & Saxe, R. (2012). Teaching replication to promote a culture of reliable science. *Perspectives on Psychological Science*, **7**, 600–604.

Fuchs, H. M., Jenny, M., & Fiedler, S. (2012). Psychologists are open to change, yet wary of rules. *Perspectives on Psychological Science*, **7**, 639–642.

Funder, D. C., Levine, J. M., Mackie, D. M., Morf, C. C., Sansone, C., Vazire, S., & West, S. G. (2014). Improving the dependability of research in personality and social psychology: Recommendations for research and educational practice. *Personality and Social Psychology Review*, **18**, 3–12.

Gilbert, E. (2014, May). Reproducibility project: Results. In E. B. A. Nosek (Chair), *The reproducibility project: Estimating the reproducibility of psychological science*, Symposium conducted at the meeting of Association for Psychological Science, San Francisco, CA.

Giner-Sorolla, R. (2012). Science or art? How aesthetic standards grease the way through the publication bottleneck but undermine science. *Perspectives on Psychological Science*, **7**, 562–571.

Gray, K., & Wegner, D. M. (2013). Six guidelines for interesting research. *Perspectives on Psychological Science*, **8**, 549–553.

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin*, **82**, 1–20.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, **2**(8), e124.

Ioannidis, J. P. A. (2008). Why most discovered true associations are inflated. *Epidemiology*, **19**, 640–648.

Ioannidis, J. P. A. (2012). Why science is not necessarily self-correcting. *Perspectives on Psychological Science*, **7**(6), 645–654.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, **23**, 524–532.

Kaiser, C. R. (2012). Campaign for real data. *Dialogue*, **26**, 8–10.

King, L.A. (2012). Science: A true story. *Dialogue*, **26**, 6–8.

Koole, S. L., & Lakens, D. (2012). Rewarding replications: A sure and simple way to improve psychological science. *Perspectives on Psychological Science*, **7**, 608–614.

Lakens, D., & Evers, E. R. K. (2014). Sailing from the seas of chaos into the corridor of stability. *Perspectives on Psychological Science*, **9**(3), 278–292.

Ledgerwood, A. (2014). Introducing the special section on advancing our methods and practices. *Perspectives on Psychological Science*, **9**(3), 275–277.

Ledgerwood, A., & Sherman, J. W. (2012). Short, sweet, and problematic? The rise of the short report in psychological science. *Perspectives on Psychological Science*, **7**, 60–66.

Maner, J. K. (2014). Let's put our money where our mouth is: If authors are to change their ways, reviewers (and editors) must change with them. *Perspectives on Psychological Science*, **9**, 343–351.

Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. *Psychological Methods*, **9**, 147–163.

Miguel, E., Camerer, C., Casey, K., Cohen, J., Esterling, K. M., Gerber, A., … Van der Laan, M. (2014). Promoting transparency in social science research. *Science*, **343**(6166), 30–31.

Mitchell, J. (2014, July). "On the emptiness of failed replications" [Web log comment]. Retrieved from http://wjh.harvard.edu/~jmitchel/writing/failed_science.htm#_ edn5

Monroe, B. M. (2014). The trouble with social psychology. *Manuscript in preparation*.

Nelson, L. D., Simmons, J. P., & Simonsohn, U. (2012). Let's publish fewer papers. *Psychological Inquiry*, **23**, 291–293.

Neuroskeptic (2014, July). On "On the emptiness of failed replications" [Web log comment]. Retrieved from http://blogs.discovermagazine.com/neuroskeptic/ 2014/07/07/emptiness-failed-replications/#.VKRM42TF95z

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods*, **5**, 241–301.

Nosek, B. A., & Bar-Anan, Y. (2012). Scientific utopia: I. Opening scientific communication. *Psychological Inquiry*, **23**, 217–243.

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science*, **7**, 615–631.

Petty, R. E., Fleming, M. A., & Fabrigar, L. R. (1999). The review process at PSPB: Correlates of interreviewer agreement and manuscript acceptance. *Personality and Social Psychology Bulletin*, **25**, 188–203.

Popper, K. (1963). *Conjectures and Refutations: The Growth of Scientific Knowledge*. Oxford, UK: Routledge & Kegan Paul.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, **86**, 638–641.

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

Schimmack, U. (2012). The ironic effect of significant results on the credibility of multiple-study articles. *Psychological Methods*, **17**, 551–566.

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

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, **22**, 1359–1366.

Simmons, J. P. (2014, April 4). MTurk vs. the lab: Either way we need big samples. [Web log comment]. Retrieved from http://datacolada.org/author/joe/

Simons, D. J. (2014). The value of direct replication. *Perspectives on Psychological Science*, **9**, 76–80.

Smart, J. J. C. (1963). *Philosophy and Scientific Realism*. London, UK: Routledge & Kegan Paul.

Spellman, B. A. (2012). Scientific utopia … or too much information? Comment on Nosek and Bar-Anan. *Psychological Inquiry*, **23**, 303–304.

Srivastava, S. (2012). Groundbreaking or definitive: Journals need to pick one. *Dialogue*, **26**, 10–12. 2011/12/31/groundbreaking-or-definitive-journals-need-to-pick-one/

Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance-or vice versa. *Journal of the American Statistical Association*, **54**, 30–34.

Sterling, T. D., Rosenbaum, W. L., & Weinkam, J. J. (1995). Publication decisions revisited: The effect of the outcome of statistical tests on the decision to publish and vice versa. *The American Statistician*, **49**, 108–112.

Stroebe, W., & Strack, F. (2014). The alleged crisis and the illusion of exact replication. *Perspectives on Psychological Science*, **9**, 59–71.

Stroebe, W., Postmes, T., & Spears, R. (2012). Scientific misconduct and the myth of self-correction in science. *Perspectives on Psychological Science*, **7**, 670–688.

van Fraassen, B. C. (1980). *The Scientific Image*. Oxford, UK: Oxford University Press.

Vazire, S. (2014, June). Another $*%#! Blog Post about Repligate* [Web log comment]. Retrieved from http://sometimesimwrong.typepad.com/wrong/2014/06/another-blog-post-about-repligate.html

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L. J., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science*, **7**, 632–638.

Young, N. S., Ioannidis, J. P. A., & Al-Ubaydli, O. (2008). Why current publication practices may distort science. *PLoS Medicine*, **5**(10), e201.