Marketing Letters
https://doi.org/10.1007/s11002-023-09680-9

Understanding effect sizes in consumer psychology

Rodrigo S. Dias[1] · Stephen A. Spiller[2] · Gavan J. Fitzsimons[3]

Accepted: 19 April 2023

Abstract

Over the past decade, behavioral scientists have learned that many findings in the field may not replicate, leading to calls for change in how behavioral research is conducted. Krefeld-Schwalb and Scheibehenne (2023) examine changes in the methodological practices in consumer research between 2008 and 2020. They find that sample sizes have increased and that effect sizes have decreased. In this article, we take these findings as a starting point and reflect on how we can further improve methodological practices in the field. We argue that in order to build a more replicable, rigorous field, we must place effect sizes at the center of scientific reasoning. Specifically, we make four claims about effect sizes that we hope will help consumer researchers plan, conduct, and interpret their research: (1) effect sizes in consumer psychology are small, and that is a natural consequence of the field's maturity; (2) effect sizes need to be contextualized; (3) our samples are still too small to detect the small effects of modern empirical consumer research; and (4) larger samples do not inherently generate smaller effects. It is our hope that the current article increases the field's understanding about effect sizes and motivates researchers to place effect sizes at the center of their scientific reasoning. By thinking carefully about effect sizes, we believe we can collectively improve methodological practices and confidence in the findings of consumer psychology.

Keywords Research methods · Effect sizes · Sample sizes · Replication · Open science

✉ Rodrigo S. Dias
rodrigo.dias@duke.edu

Stephen A. Spiller
stephen.spiller@anderson.ucla.edu

Gavan J. Fitzsimons
gavan@duke.edu

[1] Fuqua School of Business at Duke University, 100 Fuqua Drive, Durham, NC 27708, USA

[2] Marketing and Behavioral Decision Making, UCLA Anderson School of Management, Los Angeles, USA

[3] Marketing and Psychology, Duke University, Durham, USA

Published online: 29 April 2023

1 Introduction

A series of events over the past decade gave rise to what became known as the replication crisis (Nelson, Simmons, and Simonsohn 2018). The realization that many findings in the behavioral sciences may not replicate sparked calls for change in how behavioral research is conducted. In their article, Krefeld-Schwalb and Scheibehenne (2023) investigate changes in the methodological practices in consumer research between 2008 and 2020 in three of the field's premier journals (*Journal of Consumer Research*, *Journal of Marketing Research*, and *Journal of Consumer Psychology*). They find that the number of studies per paper doubled, the median sample size per cell increased from 48 to 104 participants, and the median effect size decreased from $r = .25$ to $r = .16$. We agree with their conclusion that the increase in sample size is a positive development for the field, and we believe that there is room for further improvement.

This article aims to put Krefeld-Schwalb's and Scheibehenne's (2023) findings in perspective. We take a bird's eye view of experimental consumer research and ask: How can we further improve the methodological practices in the field toward the goal of building a more replicable, rigorous discipline? We believe that part of the answer lies in placing effect sizes at the center of scientific reasoning. In the pages that follow, we describe four key points about effect sizes that can assist researchers plan, conduct, and interpret their research. Our hope is that the current article motivates researchers to think carefully about effect sizes and that careful thinking about effect sizes increases methodological rigor in the field.

2 What is an effect size, and why is it useful?

What is an effect size? At a conceptual level, an effect size is a quantitative measure of the magnitude of the relationship between two variables. Effect sizes may come in various forms (e.g., difference between means, variance explained, raw or standardized), but, in essence, they are simply a representation of the strength of the association between two variables. Despite this simple definition, researchers often conflate effect sizes with what they are not. For example, it is not uncommon for researchers to state that an effect is large, noting a small *p*-value or a seemingly large difference in the height of two bars. While *p*-values and the look of a bar graph depend on the size of an effect, they also depend on other factors (e.g., the sample size, the scale of the axis). Thus, neither *p*-values nor a graph, in and of themselves, accurately or directly represent the size of an effect.

Why is it useful to think in terms of effect sizes? There are three main reasons why it is useful to think in terms of effect sizes. First, effect sizes provide a common language to compare the magnitude of different results. An effect that explains 2% of the variance of an outcome explains twice the variance of an effect that explains only 1% of the variance of that same outcome in that context; a manipulation whose effect was $d = 0.60$ was three times as large of a difference between groups as a manipulation whose effect was $d = 0.20$. The same direct comparison cannot be made with *p*-values. For example, an effect with a *p*-value of .015 may be larger, smaller, or similar in size to an effect with a *p*-value of .030. In addition to allowing for a direct comparison of different effects, thinking in terms of effect sizes permits the researcher to examine the practical significance of their results. While different effect sizes within a given context or paradigm can be directly compared to one another, their meaning and practical significance can only be assessed with respect to the broader context of the phenomenon being studied. A final benefit of thinking carefully about effect sizes is that it allows for better planning of sample size, a point we return to later in this article.

3 Four things consumer researchers should know about effect sizes

3.1 Effect sizes in consumer research are small

Consumer research is marked by two defining features. First, most consumer-relevant phenomena are multiply determined. That is, consumers' judgments, decisions, and behaviors can rarely be traced back to a single causal factor. Second, because effects that can be easily observed in the absence of empirical data are unlikely to make their way to the field's premier journals, we typically study effects that are not plainly obvious. To the extent that we study non-obvious factors that affect multiply-determined phenomena, expecting effect sizes in consumer psychology to be large is unrealistic. Studying small effects is the very nature of what we do.

It is important to note that this argument pertains to what consumer psychologists typically study today—some 60 plus years into the field's evolution—and not to consumer behavior itself; many factors clearly have a large effect on consumer behavior (e.g., prices, default options). Researchers have moved their attention toward less obvious drivers of behavior in part because easier-to-observe factors have already been documented in the early days of the field. Critically, we are not suggesting that studying small effects is detrimental to scientific progress in the field; studying small effects is a natural consequence of the field's maturity.

3.2 Effect sizes need to be contextualized

Accepting that the effects we currently study are small, however, does not imply that they are unimportant or lack practical significance. While effect sizes provide a common language for comparing the magnitude of different results, the substantive significance of each effect can often only be assessed with respect to the context of the phenomenon under investigation. Thus, statistically small effects can have important implications. When judging the substantive significance of an effect, we believe there are three main questions researchers ought to consider.

3.2.1 Measurement

How does the measure used in the study translate to impact in the real world? A small effect found in an arbitrary metric (e.g., willingness to pay on a 7-point scale) may reflect an important effect when translated into a real-world situation (e.g., an actual dollar amount a consumer will pay for a purchase). Conversely, it is also possible that a seemingly large effect observed in a study does not translate into meaningful real-world impact.

3.2.2 Longitudinal dynamics

Might the effect observed in the study accumulate or dissipate over time? An effect may have a small effect on consumers at a single point in time, but it may carry over to later periods and accumulate. For example, a researcher may find that a particular intervention increases retirement savings by a small amount. Over time, however, this small effect may accumulate, culminating in a meaningful impact on the total amount saved by consumers.

3.2.3 Selection

Is the effect in a specific sample similar to the effect in the target population? Oftentimes, the effect observed in a particular sample may be different from the effect one would observe in the population of interest. For example, a researcher studying factors affecting intention to save for retirement may encounter smaller effect sizes when relying on a sample that does not perfectly reflect the target population. For example, a sample of students may yield smaller effects than a sample drawn from the target population of workers saving for retirement. Thus, depending on the nature of the sample selection, a study may overstate or understate the size of an effect in the population of interest.

3.3 Our samples are too small to detect small effects

Putting effect sizes at the center of the scientific reasoning invites researchers to think about whether their samples are large enough to detect the effect they are studying. We have suggested that effect sizes in prestigious consumer research outlets are likely to be small in the modern era. But how small? Krefeld-Schwalb and Scheibehenne (2023) estimate that the median effect size in published consumer research in 2020 was $r = .16$. Due to publication bias (Sterling, Rosenbaum, and Weinkam 1995) and questionable research practices (John, Loewenstein, and Prelec 2012), this estimate is likely an upper bound of the median effect size one can expect in the type of consumer research being conducted and published in top outlets in 2023. Indeed, recent research has suggested that effect size estimates based only on published work may be inflated by as much as a factor of two. For instance, a collection of large-scale, multi-site replications yielded effect sizes that were, on average, half the magnitude of the effect sizes in the original studies (Open Science Collaboration 2015). In another investigation, effect sizes from pre-registered studies were 55% smaller than effect sizes in a random sample of non-pre-registered studies (Schäfer and Schwarz 2019).

In this section, we ask: how big of a sample do consumer researchers need in order to have adequate statistical power to detect an effect? We took the estimate of the median effect size in consumer research provided by Krefeld-Schwalb and Scheibehenne (2023) as a starting point (r = .16) and adjusted it downwards given three different levels of bias (15%, r = .136; 30%, r = .112; and 45%, r = .088). We then calculated the sample size required to detect these four effect sizes in a 2-cell between-subjects design given an alpha level of 5%, as a function of the desired statistical power. Figure 1 plots the results.

This figure plots the total sample size required in a 2-cell between-subjects design as a function of the desired statistical power, for four different population effect sizes. Each population effect size is based on the median effect size estimate provided by Krefeld-Schwalb and Scheibehenne (2023), assuming either no bias or a bias of 15%, 30%, or 45%. The vertical light blue line represents x = .80

How much statistical power would a researcher have to detect an effect similar in size to the median true effect in the field, relying on a sample of the same size as the median sample in consumer research in 2020 (104 participants per cell)? The results shown in Fig. 1 paint a clear picture. In the best-case scenario (i.e., assuming that the published literature provides an unbiased estimate of the typical effect size), 64%. In the worst-case scenario (i.e., assuming a 45% bias in the published literature), 25%. In either case, 104 participants per cell do not afford consumer researchers adequate statistical power to detect the typical effect in the field. Despite a two-fold increase in sample sizes from 2008 to 2020, our studies are still underpowered, and our samples are too small.

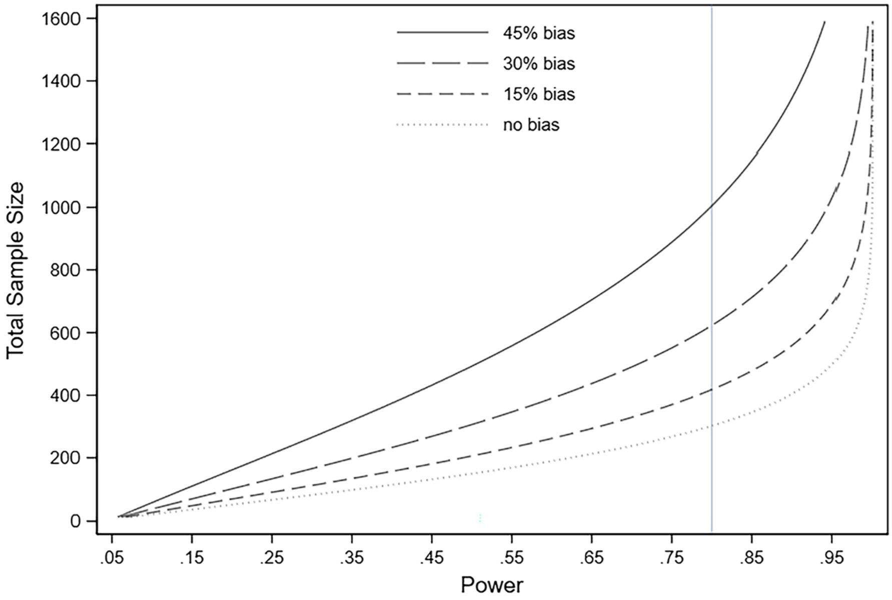

Fig. 1 Sample size required to detect the median effect size found by Krefeld-Schwalb and Scheibehenne (2023) in a 2-cell between-subjects design, assuming no bias or a bias of 15%, 30%, or 45%,as a function of the desired statistical power.

Of course, deciding the adequate level of statistical power is not an easy task, and what constitutes an underpowered study is subjective. Researchers can make well-informed decisions about statistical power by explicitly considering the relative importance of finding a false positive (a type I error, or α) and a false negative (a type II error, or β) result. Since statistical power is directly related to the probability of observing a type II error (statistical power = 1 – β), a researcher is implicitly balancing the cost of a false positive result and the cost of a false negative result when choosing a particular level of statistical power. For example, assuming a significance level α = 5%, a study with 90% power to identify an effect has a 10% chance of yielding a false negative result and a 5% chance of yielding a false positive result.

In some research contexts, a false negative result is arguably as costly as a false positive result. Consider initial tests of potential interventions: ruling out a promising candidate from consideration too soon misses out on large potential gains, whereas including an ineffective candidate at an early stage incurs the (perhaps lower) opportunity cost of further testing of the next-best candidate instead. In other research contexts, a false positive result is arguably costlier than a false negative result. Consider establishing a claim when theory-testing in a domain where a single false-positive leads a field to devote time and money on downstream consequences, potentially building a field on a weak foundation. Nonetheless, even when false positive results are extremely costly, researchers should care a great deal about type II errors (and thus statistical power) when designing studies, given that the motivation for running a study in the first place is to test for evidence of an effect. Indeed, many have noted that studies with too little power are inefficient, or even unethical (Altman 1980).

A common recommendation is to set statistical power to 80% (Cohen 1992). This implies that, given their assumptions about the population effect size, a researcher is willing to accept a false negative rate that is four times as high as the false positive rate if the population effect size is zero. We believe this is a sensible minimum threshold for most contexts within consumer psychology, and that a higher level of power may often be desired. To achieve this level of statistical power, however, consumer psychologists need to substantially increase their sample size. Even if the median effect size estimated by Krefeld-Schwalb and Scheibehenne (2023) is an unbiased estimate of the typical effect in the field, researchers would need to increase their samples to 151 participants per cell to have 80% power for a simple two-cell design. If the typical effect size in consumer psychology is 15%, 30%, or 45% smaller than the effect estimated by Krefeld-Schwalb and Scheibehenne (2023), then consumer researchers need to increase their samples to 209, 312, and 503 participants per cell to have 80% power.

3.4 Larger samples do not inherently generate smaller effects

Krefeld-Schwalb and Scheibehenne (2023) report that the median sample size per cell increased from 48 participants in 2008 to 104 in 2020, while the median effect size decreased from $r = .25$ to $r = .16$ in the same period. They find that the decrease in the median effect size over time was statistically significant and that this decrease was no longer significant after controlling for sample size, noting a strong correlation between effect sizes and sample sizes.

If a reader were to come away with the impression that smaller samples *inherently* generate larger effects, however, that would be a mistake; smaller samples simply generate more *varied* effects. To the extent that larger effects are more likely to be published (i.e., publication bias), we will observe larger effects in the published literature if we collectively rely on smaller samples. A simple thought experiment illustrates this point. Consider two groups of researchers conducting multiple studies investigating the same research question using the same research design. One group relies on samples of 50 participants per study; the other group relies on samples of 500 participants per study. Collectively, the two groups will observe the same average effect, but the first group will have more studies in which the observed effect size is much smaller and also much larger than the true effect size. If the probability of publishing a result depends on the magnitude of an effect, then the first group of researchers will, collectively, have a collection of *published* studies with larger effects.

4 Conclusion

Krefeld-Schwalb and Scheibehenne (2023) have documented an interesting shift in the methodological practices underlying the articles published in top consumer psychology outlets in the past decade. Sample sizes have increased, while effect sizes have decreased. Our goal with this article was to help ground this finding with an enhanced understanding of effect sizes. We have argued that it is neither surprising nor troubling that effect sizes in modern consumer research are small. We have also argued that the field has not gone far enough in terms of increasing the sample sizes of its consumer research investigations. We hope that the current article will help lead to even greater impact of the important analysis by Krefeld-Schwalb and Scheibehenne (2023).

Declarations

Ethical approval Not applicable.

Informed consent Not applicable.

Conflict of interest The authors declare no competing interests.

References

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), 4716.

Altman, D. (1980). Statistics and ethics in medical research. *British Medical Journal, 281*(6268), 1336–1338.

Cohen, J. (1992). Statistical power analysis. *Current Directions in Psychological Science, 112*(1), 98–101.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science, 23*(5), 524–532.

Krefeld-Schwalb, A., & Scheibehenne, B. (2023). Tighter nets for smaller fishes? Mapping the development of statistical practices in consumer research between 2008 and 2020. *Marketing Letters.*

Nelson, L. D., Simmons, J., & Simonsohn, U. (2018). Psychology's renaissance. *Annual Review of Psychology, 69*, 511–534.

Schäfer, T., & Schwarz, M. A. (2019). The meaningfulness of effect sizes in psychological research : differences between sub-disciplines and the impact of potential biases. *Frontiers in Psychology, 10*(April), 1–13.

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*(1), 108–112.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.