Psychology. Journal of the Higher School of Economics.
2020. Vol. 17. N 4. P. 803–811. DOI: 10.17323/1813-8918-2020-4-803-811

*Discussions*

# DO EFFECT SIZES IN PSYCHOLOGY LABORATORY EXPERIMENTS MEAN ANYTHING IN REALITY?

**R.F. BAUMEISTER**[a]

[a] *The University of Queensland, Brisbane QLD 4072 Australia*

**Abstract**

The artificial environment of a psychological laboratory experiment offers an excellent method for testing whether a causal relationship exists, — but it is mostly useless for predicting the size and power of such effects in normal life. In comparison with effects out in the world, laboratory effects are often artificially large, because the laboratory situation is set up precisely to capture this effect, with extraneous factors screened out. Equally problematic, laboratory effects are often artificially small, given practical and ethical constraints that make laboratory situations watered-down echoes of what happens in life. Furthermore, in many cases the very notion of a true effect size (as if it were constant across different manipulations and dependent variables) is absurd. These problems are illustrated with examples from the author's own research programs. It is also revealing that experimental effect sizes, though often quite precisely calculated and proudly integrated into meta-analyses, have attracted almost zero attention in terms of substantive theory about human mental processes and behavior. At best, effect sizes from laboratory experiments provide information that could help other researchers to design their experiments, — but that means effect sizes are shop talk, not information about reality. It is recommended that researchers shift toward a more realistic appreciation of how little can be learned about human mind and behavior from effect sizes in laboratory studies.

**Keywords:** effect size, experiment, laboratory, statistical interpretation.

Science relies heavily on quantification. Increasingly precise measurement is a hallmark of scientific progress. Scientists seek to measure important things, and the association between precise measurement and scientific importance furnishes a heuristic assumption that what is measured with great precision must be highly important.

The heuristic association between importance and precision can be misleading, however. Quantity of life can be measured much better than quality, and so public policy has focused much more on improving quantity than quality of life. When I ask people whether they would relinquish all pleasures in order to live 300 years, they laugh and say no, but each time some pleasure is linked to shorter lifespan, people start to eliminate it from their lives.

Half a century ago, psychologists emphasized significance testing as a binary judgment as to whether an experimental result was significant or not. The purpose of an experiment was to establish whether a causal relationship existed between the independent and dependent variables. In that context, the purpose of significance testing was to avoid being wrong too often. Recently, however, psychological scientists have shifted toward

estimating the effect sizes of their laboratory findings. Meta-analyses combine the results of many studies to proudly assert they have discovered the effect size, often computed precisely to several decimal places. Reporting of effect sizes has become standard practice in many journals, and some editors have gone so far as to abandon and prohibit null hypothesis significance testing, instructing authors to report only the size of the effects in their samples.

But do those effect sizes mean anything? They may serve as guidelines for other laboratory experiments. Crucially, however, the psychological laboratory is an artificial environment. Outside the laboratory, sizes of effects do matter: Saving ten thousand lives is better than saving ten lives. Inside the laboratory, effect sizes would mainly seem to matter insofar as they predict real-world effects. I contend that laboratory effect sizes are not reliable predictors of real-world effect sizes, in which case they do not matter.

I realize this argument is provocative. To minimize the degree of offending other people, I shall favor examples from my own research programs.

## The Case against Effect Sizes

There are several reasons to doubt that the effect size of a laboratory experiment contains any useful information. Presumably scientists do experiments to learn about psychological processes outside the laboratory, rather than simply learning about the inner workings of their experiments. However, generalizing from a laboratory finding to the external world is hazardous – indeed, for effect sizes, probably impossible.

### *Artificially Inflated*

First off, laboratory experiments will often furnish inflated estimates of effect sizes. Researchers set up their experiments carefully to give the best chance of finding a significant result, given that there is a genuine causal relationship. Within each treatment condition, each participant experiences almost exactly the same stimuli, environment, and procedure very much unlike daily life, in which circumstances differ and no two people are likely to encounter exactly the same experience. Moreover, participants know they are in an experiment and follow instructions as to where to focus their attention and when to respond. Laboratory experiments are designed to be best-case scenarios for testing their hypotheses. Some effects may be large and robust in the laboratory but hardly ever happen outside the lab. The large and impressive body of work on social cognition is based on studying human thought under ideal conditions. Outside the laboratory, people have more diverse distractions, have not been instructed to pay attention, and may be preoccupied with other issues. Therefore, out in the world, many effects will be considerably smaller than in the lab.

As an example, Alquist et al. (2015) manipulated people's beliefs about free will by having them re-write sentences affirming or denying free will, in their own words. After that we measured how they generated counterfactual beliefs. The causal relationship was supported. On various measures, we found eta squared to vary between .06 and .11. (After denying free will, they generated fewer counterfactuals about actual events from their past.) But generalizing the effect size to what happens outside the laboratory is hazardous. After all, in everyday life people do not frequently reformulate sentences about free will,

then reflect on prior experiences and generate new lists of how things could have gone differently. Perhaps they do other, vaguely similar things. Still, there is no way of knowing whether the natural environment in people's everyday life would amplify or dilute the causal effect size observed in the laboratory.

### *Artificially Deflated*

On the other hand, many laboratory studies are constrained by ethical and practical concerns. The experience they create can only be a feeble echo of what likely happens outside the laboratory.

My studies on interpersonal rejection have relied on simple procedures such as telling research participants that no one else in the group chose to work with them, or telling them that computer analysis of their responses to a questionnaire yielded the prediction that they will end up alone in life (e.g., Twenge, Baumeister, Tice, & Stucke, 2001). These obviously pale beside the impact of being rejected by the love of one's life, or one's preferred medical school. In a similar vein, my studies on choking under pressure (Baumeister, 1984) used a manipulation in which students were offered a few dollars to perform well on a video game or hand-held game, and clearly this falls far short of the pressure of taking a major final examination or competing for a sports championship in front of a large crowd. In these cases, the size of an effect outside the laboratory will almost certainly be larger than what the lab study can find.

In some of my other work, we offer small amounts of money as incentives. Again, the amounts of money we can afford to offer in an experiment pale beside what is often at stake in decisions outside the laboratory. We can study how cash incentives change decisions, but the effects are likely smaller than how large incentives operate.

### *Unreliable Estimates*

A further argument against effect sizes is that most experiments with typical (and pragmatically viable) sample sizes simply lack the precision to furnish an effect size, even within the laboratory. Simonsohn (2014) has estimated that simply to tell whether an effect is large, medium, or small, based on the convention that the d would be .8, .5, or .2, would require three thousand participants per cell. A basic 2×2 experiment, which was the standard design in social psychology for decades, would therefore require 12,000 participants in order to justify the crude conclusion as to whether the effect size is large, medium or small, let alone any greater precision. Considering that hardly any high-involvement experiment comes close to N = 12,000, the only way to get close is with meta-analyses that combine results from dozens of studies. Moreover, variations in procedure and sampling would undermine even that attempt. That means that an experiment is not worth doing (as a way of establishing effect size) unless there are many other, *essentially identical* experiments being conducted.

### *Very Idea is Incoherent*

More broadly, the very idea of a true effect size for a psychological variable often makes no sense. Return for a moment to my studies on interpersonal rejection. What is the effect size of being rejected?

The very question of such a true size is absurd. Note, however, that there have been a great many laboratory experiments on rejection, and it is quite possible for a meta-analyst to compile a few hundred of these and compute an average effect size, though this is normally done with a specific dependent variable (e.g., state self-esteem). My argument, however, is that such an exercise would be essentially worthless and meaningless, except for advising other lab researchers how to set up their experiments through a power calculation. There are several reasons for this lack of wider value.

First, there is the question of what is the dependent variable. Some writers speak of effect sizes as if all dependent variables are interchangeable. In our studies of interpersonal rejection, we have found large effects on behavior but small to negligible effects on emotion. Even the behavioral findings should not be lumped into the same bag. We tested effects on specific behaviors based on our theorizing, and so we have looked at promising candidates (e.g., aggression, helping). Plenty of behaviors are likely unaffected by interpersonal rejection, even though we also found some large effect sizes. Thus, exactly the same manipulation can produce both large and negligible effects, with different dependent variables. Neither is the true effect size of the independent variable.

Second, the manipulation of the independent variable also is of questionable generality and may not even qualify as a single event. Any averaging of effect sizes must again take into account the problem already noted, which is that the lab rejection almost certainly lacks the power and impact of an important rejection in everyday life. (The other problem already noted may also be relevant: In the lab, people are fully engrossed in the activity and so probably pay full attention to the rejection. So it is possible that some laboratory effects of rejection are larger than what would be found in actual social life.) Divorce is a form of rejection, but so is asking a stranger in an airport lounge whether one may sit here and being told that seat is already taken. Effect sizes (even on the same dependent variable) for those two rejections are probably quite different, however.

My work on ego depletion is relevant here. Hundreds of studies in multiple laboratories have replicated the basic effect, which qualifies it as one of the most frequent findings in social psychology (see Baumeister & Vohs, 2016). Yet others have questioned its existence, sometimes because they failed to find the effect in their own laboratories. A large multi-site replication by Vohs, Schmeichel, Funder, and many colleagues (2018) showed a significant effect, which would seemingly lay to rest once and for all the objection that there is no effect. Moreover, just during the past year, two additional large-scale projects did also find a significant depletion effect on subsequent self-regulation (Garrison, Finley, & Schmeichel, 2019; Dang et al., 2019). Critics have therefore retreated to griping about the effect sizes in these studies, some of which were relatively small.

The relatively small effect sizes for multi-site replications are likely to be a regular feature. As I said, a laboratory experiment is typically set up to be a best-case situation for testing the causal hypothesis. Procedures and measures are not selected at random but are carefully chosen to be suitable for local conditions and samples. A multi-site study inevitably eliminates much of that calibration, typically using identical procedures for all sites, and so it is to be expected that multi-site replication studies will routinely produce smaller effect sizes than the original study. Again, this is a basic fact about experimentation and has no implications or relevance for understanding or predicting effects outside the laboratory.

More fundamentally, is it even meaningful to talk about the size of ego depletion laboratory effects? Ego depletion is a form of psychological fatigue, so by analogy one might as well ask, how big is the effect of being tired? Put that way, the question seems absurd, though the absurdity seems to escape many writers who continue to speak about it. But to talk about an effect size of tiredness is meaningless if one ignores the two dimensions already noted: *how* tired is the person, and effect on *what*? Would anyone think that a statement such as "a tired person will do 0.12 standard deviations worse than a non-tired one, across all tasks" is meaningful?

Regarding ego depletion, too, the former issues are relevant again. Many studies have used a five-minute manipulation, which almost certainly will produce a much smaller effect than, say, a couple hours of grueling work or inner struggle outside the laboratory. Even inside the laboratory, recent studies using longer-lasting inductions of ego depletion produced large effects on the manipulation checks (one was over 4 SD) and medium-sized effects on the dependent variable (Sjestad & Baumeister, 2018; see also Guilfoyle, Struthers, van Monsjou, & Shoikhedbrod, 2019). Other experiments have observed a similar pattern: Blain and colleagues (2016) found that a brief depletion manipulation did not produce a significant effect on future discounting, but a more severe and exhaustive task did. With relevance far beyond ego depletion and self-control research, the broader implication is that the dosage of the causal factor is likely to matter a great deal for the effects we observe in the laboratory. In my view, psychological scientists would benefit from taking this point more seriously before making strong and generalized claims about *the* effect size.

## *A Physical Analogy*

Heat is a physical variable. Imagine an engineer or physicist claiming to have discovered the effect size of heat. That would be ridiculous, and the researcher would become a laughingstock. The idea is incoherent for both reasons. That is, one would have to ask, how much heat? As well as, effect on what? It is quite possible to draw broad conclusions about the causal effects of heat, as in increasing molecular motion, expanding solid things, and melting liquids. But no serious physicist would talk about there being a true effect size of heat, independent of quantity of heat and of specific type of object. In the mountains, heat increases during the summer, but the effect is bigger on the snow than on the rocks. With enough heat, though, the rocks would melt also, thereby making a much bigger mess than the melted snow. Still, is that a small "effect size of heat", or a large one?

The psychological effects of heat are even more variable. Abundant evidence indicates that hot temperatures increase aggression. Yet some of the highest temperatures people encounter are in saunas, and these hardly ever elicit aggression. Meanwhile, heat above a certain temperature causes immediate death, so that all behavior and cognition cease — which makes heat an extremely strong moderator of all psychological processes.

## *Public Policy*

Consider a variable that psychologists study for which public policy could benefit from precise information: alcohol consumption. Does it make sense to speak of a true effect size of drinking alcohol? Given alcohol's effects on crime and traffic accidents, such

information would be welcome. But some things are not affected by alcohol. So, again, rather than speaking of an effect size of alcohol, one has to ask, how much alcohol, and effects on what?

Indeed, the policymakers already seem to recognize distinctions that some laboratory researchers who focus on effect sizes do not. In most Western countries, in contrast, the amount one has consumed is important. Drivers are permitted to have had a small amount of alcohol but not a large amount. In fact, the amount is not even usually the issue, in recognition of the fact that two beers will have a much bigger effect on a 95-pound woman than a 300-pound man, and they will also have a bigger effect on someone with an empty stomach than someone who consumed them alongside a giant Wiener schnitzel and a large plate of fried potatoes. Alcohol and driving laws tend to set the limit in terms of the level of alcohol in the blood. Although it is fair and correct to say that drinking alcohol impairs driving, it is incoherent to claim there is a specific effect size, even on driving. If one can specify how much alcohol, in proportion to body weight and controlling for other factors, a precise estimate of how much worse one drives may become somewhat more plausible but still remains elusive. Moreover, it is doubtful that findings from laboratory experiments could yield an effect size estimate that is a reliable guide to what happens to real drivers out on the highways.

## Conclusion

The fact that something can be computed from laboratory observations does not entail that it offers meaningful information about human everyday life. The unthinking emphasis on effect sizes should be replaced by careful, thoughtful evaluation of what those numbers mean. For real-world data, with proper appreciation of context, effect sizes are quite important, and I can understand why many researchers and editors find them more important and informative than significance testing. Applied researchers in particular already pay much more attention to effect sizes than significance testing. And with good reason: The applied goal is to improve society, and large improvements are better than small ones. But in a laboratory experiment, the size of the effect bears no reliable relationship to what happens in daily life, except perhaps when daily life includes many situations that very precisely resemble the laboratory setting.

While writing this paper and discussing the issue over the past year, I have only come across one suggestion for how lab effect sizes could furnish potentially useful information about reality. That would arise if one effect is reliably larger than another. As noted above, our work on rejection has consistently found larger effects on behavior than on emotion. Blackhart et al.'s (2009) meta-analysis concluded that the effects of being rejected are generally larger than the effects of being socially accepted, when both are compared to a common neutral control condition. This difference corresponds well to the general pattern that bad things are stronger than good ones, which some of us concluded in a review article (that, ironically, did not use meta-analysis or discuss specific effect sizes; Baumeister, Bratslavsky, Finkenauer, & Vohs, 2001). The rejection difference was already evident in significance tests, given that in many studies the acceptance condition does not differ from the neutral control, whereas the rejection condition does. Nevertheless, I can imagine drawing a potentially useful conclusion about reality if there are two frequently

significant effects, of which one is consistently larger than the other. To be sure, that assumes no moderators between lab and world that cause a crossover interaction, and that assumption could be wrong. Still, it seems reasonable to start with the default assumption that a *reliably and consistently* larger lab effect is also larger out in the world, when the lab study compares them under identical conditions and with the same measures.

Laboratory effect sizes can provide useful technical information for other researchers who intend to use the same manipulations and measures. The others can use that information as input to power calculations for choosing their own sample size. But then, even in the best case, laboratory effect sizes are merely a form of shop talk rather than building theory or furnishing information useful outside the laboratory. Plus, again, thousands of participants per cell are needed to estimate of whether the effect is large, medium, or small.

Social psychology in particular is emerging from a period of troubled soul-searching and has apparently determined that its future lies in abandoning labor-intensive behavioral observation in favor of collecting large samples using Mechanical Turk workers and other online samples, who sit at computers and furnish ratings and judgments while imagining various scenarios (Anderson et al., 2019; Dolinski, 2018; Sassenberg, 2018). While such findings are of indisputable value in testing some hypotheses about people's mental states, generalizing their effect sizes to actual behavior out in the world seems quixotic if not utterly frivolous.

My opinion remains that in the social sciences, all methods have drawbacks, and so for best progress we need to use all different methods. That approach requires us to maintain a careful awareness of the limitations of each approach. The laboratory experiment remains the best method in all the social sciences for getting evidence about whether a causal relationship actually exists between some independent and dependent variables. Its usefulness for other things, including estimating what the strength of that relationship is outside the laboratory, is dubious.

## References

Alquist, J. L., Ainsworth, S. E., Baumeister, R. F., Daly, M., & Stillman, T. F. (2015). The making of might-have-beens: Effects of free will belief on counterfactual thinking. *Personality and Social Psychology Bulletin, 41*, 268–283.

Anderson, C. A., Allen, J. J., Plante, C., Quigley-McBride, A., Lovett, A., & Rokkum, J. N. (2019). The MTurkification of social and personality psychology. *Personality and Social Psychology Bulletin, 45*(6), 842–850. doi:10.1177/0146167218798821

Baumeister, R. F. (1984). Choking under pressure: Self-consciousness and paradoxical effects of incentives on skillful performance. *Journal of Personality and Social Psychology, 46*, 610–620.

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is stronger than good. *Review of General Psychology, 5*, 323–370.

Baumeister, R. F., & Vohs, K. D. (2016). Strength model of self-regulation as limited resource: Assessment, controversies, update. *Advances in Experimental Social Psychology, 54*, 67–127.

Blackhart, G. C., Nelson, B. C., Knowles, M. L., & Baumeister, R. F. (2009). Rejection elicits emotional reactions but neither causes immediate distress nor lowers self-esteem: A meta-analytic review of 192 studies on social exclusion. *Personality and Social Psychology Review, 13*, 269–309.

Blain, B., Hollard, G., & Pesiglione, M. (2016). Neural mechanisms underlying the impact of daylong cognitive work on economic decisions. *Proceedings of the National Academy of Sciences (PNAS), 113*, 6967–6972.

Dang, J., Berkman, E. T., ... Zinkernagel, A. (in press). Multi-Lab replication reveals a small but significant ego depletion effect. *Social Psychology and Personality Science*. doi:10.31234/osf.io/cjgru

Dolinski, D. (2018). Is psychology still a science of behavior? *Social Psychology Bulletin, 13*(2), Art. e25025.

Garrison, K. E., Finley, A. J., & Schmeichel, B. J. (2019). Ego depletion reduces attention control: Evidence from two high-powered preregistered experiments. *Personality and Social Psychology Bulletin, 45*, 728–739.

Guilfoyle, J. R., Struthers, C. W., van Monsjou, E., & Shoikhedbrod, A. (2019). Sorry is the hardest word to say: The role of self-control in apologizing. *Basic and Applied Social Psychology, 41*(1), 72–90. doi:10.1080/01973533.2018.1553715

Sassenberg, K. (2018). *Virtues and vices of the call for sufficient statistical power: Larger sample sizes, but lower quality methods?* Manuscript under review, University of Tübingen, Germany.

Simonsohn, U. (2014, May 1). *We cannot afford to study effect size in the lab* [Blog post]. Retrieved from http://datacolada.org/20

Sjestad, H., & Baumeister, R.F. (2018). The future and the will: Planning requires self-control, and ego depletion leads to planning aversion. *Journal of Experimental Social Psychology, 76*, 127–141.

Twenge, J. M., Baumeister, R. F., Tice, D. M., & Stucke, T. S. (2001). If you can't join them, beat them: Effects of social exclusion on aggressive behavior. *Journal of Personality and Social Psychology, 81*, 1058–1069.

Vohs, K. D., Schmeichel, B. J., & Funder, D. C. (2018, February). *A pre-registered depletion replication project: The paradigmatic replication approach*. Presented to the Society for Personality and Social Psychology, Atlanta, GA.

**Roy F. Baumeister** — Professor of Psychology, University of Queensland (Australia), Ph.D.
Research Area: self and identity, self-regulation, sexuality and gender, aggression, self-esteem, meaning, consciousness, free will, and self-presentation.
Email: r.baumeister@psy.uq.edu.au

# Размер эффекта в психологических лабораторных экспериментах: означает ли он что-то в реальности?

**Р.Ф. Баумайстер**[a]

[a] *Университет Квинсленда, Австралия, Brisbane QLD 4072 Australia*

## Резюме

Искусственная среда лабораторных экспериментов в психологии предлагает отличный способ проверки наличия причинно-следственных связей, но это обычно бесполезно для предсказания размера и мощности таких эффектов в нормальной жизни. По сравнению с эффектами в реальном мире, лабораторные эффекты часто искусственно преувеличены, потому что лабораторная ситуация организуется специально для того, чтобы ухватить именно этот эффект, а посторонние факторы отсеиваются. Другая проблема состоит в том, что лабораторные эффекты часто искусственно уменьшены из-за практических и этических ограничений, которые делают лабораторные ситуации лишь ослабленным эхо того, что происходит в жизни. Более того, во многих случаях абсурдно само понятие истинного размера эффекта, как будто он константен по отношению к различным манипуляциям и зависимым переменным. Эти проблемы иллюстрируются примерами из собственных исследовательских проектов автора. Обнаруживается также, что, хотя размеры эффекта в экспериментах часто высчитываются с большой точностью и гордо включаются в метаанализы, они почти не привлекают внимания при построении фундаментальных теорий психических процессов и поведения. Размер эффекта в лабораторных экспериментах в лучшем случае дает информацию, которая полезна другим исследователям для планирования их экспериментов, но это значит, что размеры эффекта хороши для светской беседы, а не в качестве информации о реальности. Исследователям сто́ит более реалистично оценивать пользу, которую могут принести вычисления размера эффекта в лабораторных исследованиях для изучения психики и поведения человека.

**Ключевые слова:** размер эффекта, эксперимент, лаборатория, статистическая интерпретация.

**Рой Ф. Баумайстер** — профессор, Университет Квинсленда (Австралия), доктор психологических наук.
Сфера научных интересов: личность и идентичность, саморегуляция, сексуальность и гендер, агрессия, самооценка, смысл, сознание, свобода воли, самопрезентация.
Контакты: r.baumeister@psy.uq.edu.au