RESEARCH ARTICLE

Replication and Reproducibility in Psychological, Medical and other Sciences

Authors
Professor Seppo E. Iso-Ahola
University of Maryland
School of Public Health
Department of Kinesiology

Correspondence address:
Email: Isoahol@umd.edu

Abstract

As there are no universal constants in psychological, medical and economic sciences, only constructive-phenomenon replications are meaningful. Yet, psychologists continue to perform direct replications, as evidenced by recent preregistered multilab attempts at exact replications of the ego depletion effect. Statistics are driving the replication movement into a ditch because of an overemphasis on the determination of statistical magnitude of effects while ignoring *commonsense magnitude* and other criteria for evaluating phenomena's validity, reliability, and viability. The nature of the human mind and the variability of psychological phenomena pose difficult challenges for the scientific method and insurmountable obstacles for precise replications in psychological sciences. The situation is no better in medical and economic sciences. The interaction effect of person (genetics) and environment (lifestyle) calls for constructive replications to determine, for example, drugs' efficacy as a function of group and individual differences. The vaccine-vaccination paradox is an interesting case because psychological and medical sciences meet at this intersection. In all fields, science advances by theory building and model expansion, not by replication tests of statistical hypotheses. Rigorous logical and theoretical analysis always precedes and guides good empirical tests. The nonexistence of an effect is not viable if it can withstand rigorous logical and theoretical analyses. Empirical studies are mainly evaluated for their theoretical relevance and importance, not their success or failure to exactly reproduce the original findings.

Introduction

Aims and Scope. This paper examines the role of replication and reproducibility in the establishment of scientific truth. It first explores the question whether psychological phenomena, and those in other sciences, are constant or stable enough to be reproduced. Since the original study has already confirmed the existence of a phenomenon under certain conditions, replications are placed in an arbitrary position to declare either the phenomenon's existence or nonexistence, and hence a statistically dichotomous decision about the significance or non-significance of an effect. But does science have tools to declare nothingness or that something does not exist? The paper aims to answer this question and clarify the interdependent relationship between theory and empirical data. Special attention is paid to preregistered multilab replications and their assumed superiority at determining the truth value of theoretical constructs and empirical findings. The paper further addresses the central role of statistical analysis, especially the magnitude-of-effect determination, in decision making regarding the viability of reported effects, and whether the heavy reliance on categorical statistical decisions (yes-no) is more confusing than elucidating for understanding phenomena's replicability in psychological and other sciences.

Challenges for the scientific method. If an idea is to discover and demonstrate empirically the existence of permanent laws or constants, then psychology is by far the hardest science. Unlike those in physics (e.g., speed of the light), psychological particles do not exist and thus cannot be replicated on researchers' command. Even physicists (1) agree that consciousness is the hardest problem in all of science. Whether due to its conscious or nonconscious operations, the human mind does not produce universal and unchangeable effects on and of feelings, cognitions or behaviors. A big challenge for psychological science, therefore, is that affect, cognition and behavior, and their effects, are both stable and variable, subtle and strong in different contexts and at different times, and even at the same time (2). For example, individual thinking can be simple and complex at different times and even at the same time depending on internally generated or externally imposed stimuli in different contexts.

This variable nature of the human mind poses difficult challenges for the scientific method and insurmountable problems for exact replications of psychological phenomena. However, it does not mean that psychological phenomena do not exist, only that the establishment of their boundary conditions is necessary, hard and time-consuming (5) Unfortunately, the idea of replication has been misinterpreted in psychological science to mean that phenomena either exist or they do not. This yes-no interpretation has been adopted from the physics model of replication, where the accepted existence of various phenomena (e.g., gravity waves as predicted by Einstein's theory) critically depends on successful precise replications.

The idea of replication in psychological science is based on an ill-conceived premise that humans are robot-like. Accordingly, their feelings, cognitions and behaviors are stable and constant across situations and times, and the related phenomena are therefore expected to replicate empirically. If not, failures to replicate mean that phenomena do not exist. This shortsightedness has led to the abandonment of many well-established psychological phenomena, from loss aversion to delay of gratification. Yet, exact or direct replications are never possible in psychological science because identical methodological conditions cannot be reconstructed in replication studies and because of the inherent variability of psychological phenomena due to the nature of the human mind. All of this means that replication failures are inevitable and direct

replications nothing but exercises in futility. The end result is that logically and theoretically well-developed phenomena will stand and cannot be argued not to exist on the basis of so-called failed replications (e.g., 3,4,5,6,7). Scientific inferences go far beyond statistical inferences (51,53).

Nothingness

Physicists have asked, "Why is there something rather than nothing?" In a similar vein, psychologists could ask, why do psychological phenomena exist? A simple answer is that they exist because people are psychological entities or beings. Insofar as people are human beings (and not robots), there is no such thing as nothingness in human affect, cognition and behavior. Instead, psychological phenomena exist in many forms, contexts and times and cannot be cast by empirical research into a trash bin of nothingness. Empirical science does not have tools to declare nothingness or prove negatives (e.g., God does not exist) as permanent constants or laws because:

1. The presence of evidence for **X** does not necessarily equal to the absence of evidence for **Y**, nor is **X**'s existence a precondition for **Y**'s nonexistence. More generally, the absence of evidence is not evidence for the absence of a scientific truth (23).

2. Empiricists can never test all the conditions and groups of people on earth; they cannot even think of all conditions that could give a rise to a phenomenon. Moreover, as "there is an infinite number of ideas and ways of testing phenomena, no idea ever achieves the status of final truth" (24). The scientific method can only produce "temporary winners", provisional, propositional, conditional and relative evidence, but no absolute truth (5).

3. Empiricists do not have perfectly reliable and valid measures, thus lacking "specificity" to avoid false positives and "sensitivity" to avoid false negatives, as the world has witnessed regarding the Covid-19 tests. Empirical studies also suffer from sampling errors. Even the champion of the falsifiability principle of science (25) acknowledged that theories cannot conclusively be disconfirmed because of the unreliability of experimental findings. So, for psychometric reasons alone, any empirical study is inclined to fail to replicate the original findings, and the null hypothesis is therefore never true (26).

4. All effects are "interaction" effects even though laboratories often test "main" effects (e.g., the depletion vs. the non-depletion condition). Research suggests that real-life effects are often stronger and thus potentially more replicable than the same effects obtained from artificial lab experiments (28,34,35,48). It was recently found that real-life social interaction correlated positively with replication success while none of the replication failures involved ongoing social interaction but instead, brief finger-pressing tasks on computer screens (35). For the most part, human behaviors are multi-causal in real life, and even in lab experiments. This means that the manipulation of a focal independent variable affects other causal factors (5). Besides, researchers cannot control for all possible confounding factors, or even think of all of them, as "everything is correlated with everything else, more or less" (18). Thus, a theoretically and logically developed and justified phenomenon cannot be zero empirically, and the null hypothesis cannot ever be proven by empirical data.

5. The operations of the human mind and associated behaviors are both stable and variable, predictable and elusive, unbearably simple at one time and irreducibly complex at another, and sometimes both at the same time. Thus, human thoughts are not completely reproducible from situation to situation, making replication failures inevitable. There is not a single replication

study that would have shown that participants' feelings and cognitions—both conscious and nonconscious—are the same with those of the original study participants (5). Of course, there are patterns in human thinking and behaviors, but patterns are just that, transient patterns, not permanent laws. If people fail at self-control (e.g., using profanities in public), it does not mean that self-control as a psychological phenomenon does not exist. It then follows that those psychological phenomena, unlike many (but not all) phenomena in physics, are not fixed constants in space and time. There is no cognitive dissonance particle that could irrevocably be verified and confirmed by empirical data and declared a universal constant.

To better understand stability and variability and subtlety and strength of psychological effects, it is useful to think of how stimuli are delivered in psychological experiments. Consider the following simple example of testing the effect of aggressive stimuli. In the first condition, upon entrance, participants are hit in the nose. Naturally, without individual differences, most, if not all, would react equally violently. In the second condition, participants are immediately insulted verbally. It is likely that this stimulus would produce weaker effects and more individual differences in reactions than the first stimulus. In the third condition, participants would be told an aggressive joke, and its effect would be expected to be even weaker than that of the second condition. As stimuli become subtler, so do their statistical effects in magnitude.

It has been proposed that psychology is a science of subtleties in human affect, cognition and behavior (2). The subtler the effects, the lower the Cohen's effect sizes. As an example, social priming (28,34,37,46,47) typically deals with subtle effects, but nevertheless there are over 100 successful replications of such effects in the literature (35). Thus, it is easy to fall in a trap of declaring that a given effect does not exist based on Cohen's effect size when in reality the effect, despite being small statistically, can be powerful in real life. Sarcastic comments are subtle but potentially strong and meaningful.

6. Variability in human behavior is a blessing rather than a curse. Experimental evidence suggests that behavioral variability confers survival benefits as the central nervous system was not created to repeat the same thing over and over, but be ready to improvise and produce variable responses when pursuing a prey or escaping from becoming a prey (36). Not only are there (a) interpersonal differences (between persons) but also (b) intrapersonal differences (within persons) and (c) interpersonal differences in intrapersonal change (27). Most psychological experiments study phenomena based on interpersonal differences or between-group differences. Even if an experimental and control group do not differ significantly (the depletion vs. non-depletion condition), and thus the claim about the phenomenon's nonexistence, the effect can still exist as an intrapersonal phenomenon. There can also be interpersonal differences in intrapersonal change, like when some children become better at cognitive and motor skills at a faster rate than others. However, most experiments in psychology are conducted in line with the first protocol while ignoring the other two, as exemplified by the Vohs et al. multilab replication study (9). In general, to better understand psychological phenomena, it would be as important to study their variability as their stability, what factors make some phenomena more stable and variable than others.

Stimulus properties are not the same for every person because people perceive and interpret stimuli differently. They process information differently not only under different conditions but even under the same conditions. The same coffee brand may taste sweeter one day than another.

7. Psychological phenomena can be experienced consciously or nonconsciously. The same phenomenon may be replicated successfully as a nonconscious but not conscious phenomenon, or vice versa. However, there are logical reasons to expect nonconsciously experienced phenomena to be more replicable because of people's general tendency to delegate conscious thoughts to nonconsciously processed operations (28,37). The more frequently thoughts and behaviors are repeated, the more automatic and nonconscious they become (19,20,21). As nonconscious thoughts are cognitively nondemanding, they are less liable to conscious interference, and thus, other things being equal, more repeatable and replicable. But this remains to be further analyzed theoretically and investigated empirically.

In short, the nonexistence of psychological phenomena is not logically viable. If a phenomenon can withstand rigorous logical analysis, it is then real. Is it logically and theoretically meaningful that X is related to Y? Unless the logic behind the X-Y relationship can be shown faulty, no failed replication can deny the possibility and plausibility of the relationship. Failed replications can only say that when this method was used, the phenomenon was not found in the investigated context, condition, and time. Rigorous logical and theoretical analyses are always more important than empirical studies because empirical testing depends on and starts from the theoretical development of a phenomenon; logical and theoretical analyses show what is worth testing. If Festinger had not discovered cognitive dissonance, there would be no cognitive dissonance to be replicated today (5). Of course, inductive and exploratory studies are done, but they are not confirmatory tests of X-Y relationships.

Statistics Are Driving The Replication Movement Into A Ditch

A replication's success or failure has exclusively been determined by statistical means, namely, by p-value and the associated dichotomous decision (statistically significant or nonsignificant) regarding the existence of an effect. There have been numerous calls for the retirement of statistical significance and its replacement by effect size, confidence interval and Bayes Factor (29, 51). Yet, subjectivity and artificiality loom large in these "new statistics" (29). For one thing, Cohen (30) arbitrarily labeled effect sizes in wide ranges, from small-moderate (0.2-0.5) to moderate-large (0.5-0.8) to large-very large (above 0.8). For another, when confidence intervals are used, a relationship between two variables is still considered "significant" provided that the lower bound of the confidence interval does not touch zero. For still another, the suddenly popular Bayesian analyses yield arbitrary ranges for the Bayes Factor to indicate "anecdotal evidence" (1-3), "substantial evidence" (3-10), "strong evidence" (10-30), and "very strong evidence" (30-100) (31). However, Kass and Raftery (32) recommended different labels and quantitative ranges for a varying degree of evidence. Notice the dilemma a researcher faces in selecting verbal descriptions for evidence when his/her data produce precise Bayes Factors of 10, 30, and 100. Bayesian analyses are also contentious because they force researchers to specify "priors" of data distributions for the null vs. alternative hypotheses and to make specific predictions from theories, and because they do not control Type I and Type II errors (33).

A major problem is that the replication movement continues to be driven by the dichotomous statistical decision making, even regarding the magnitude-of-effect determination. But it should not be forgotten that the magnitude of an effect is only one criterion by which phenomena can be evaluated and understood. It is important to

make a distinction between *statistical magnitude* and *commonsense magnitude.* The Vohs et al. (9) and other multilab replication studies of the self-depletion effect have shown what the statistical magnitude (effect size and Bayes Factor) looks like, but no study has shown what commonsense magnitude means. For it, consider activity choices people make following self-control depletion stemming from their work. (Ego depletion is a hypothesis that the use of self-control in a task A leads to reduced ability for self-control when performing a subsequent task B.). Do they hit a bar and drink excessively, or do they watch TV more than normally and avoid mundane chores they should be doing (e.g., laundry), or do they compensate by going to a gym where they do not regularly go, or do they run extra miles, or worst of all, do they become more aggressive and take their self-control depletion on others? In short, how do they cope with self-control depletion? These are just a few examples of potentially high-magnitude effects and different forms in which self-control depletion is manifested in real life when people try to cope with it, possibly greater and more meaningful effects than Cohen's effect sizes obtained from finger-pressing movements recorded in performance of 10-minute artificial experimental tasks. Remarkably, no study has investigated these real-life effects of self-control depletion.

Perhaps the best way to study the phenomenon would be to conduct surveys that would first ask participants if they have experienced self-control depletion, and if yes, then how frequently and under what conditions. Surveys could also ask people to indicate their activity choices (cognitive and behavioral) after they have been depleted and not depleted, or compare those who identify with ego depletion experiences with those who don't. The problem is obvious: the replication movement has driven researchers to labs to test the statistical magnitude of various effects in efforts to arrive at the pinpoint statistical verification through replication while ignoring other commonsense methods and ways of investigating phenomena. Activity choice and the degree of participation are just two outcome measures that can give much more useful information than statistical magnitude. These dependent variables can potentially be more revealing about the impact of ego depletion than effect sizes obtained from measurements of participants' performance on trivial and random tasks in labs.

In sum, it is questionable whether the "new statistics" are any better at determining phenomena's reliability, validity, replicability, and viability. All statistical methods are based on certain assumptions (e.g., study design, random selection and assignment of participants) about the sequence of events that lead to the reported statistics (52). A key point is that scientific inferences go far beyond statistical inferences as factors other than statistics are often more important (e.g., theoretical mechanisms, previous evidence, quality of data) (51,53).

Misguided Replications

To better understand how preregistered multilab replications in psychology have gone astray, it is worth taking a closer look at one recent effort designed to determine, once and for all, whether ego-depletion is a real phenomenon. As this examination of Vohs et al.'s multilab study (9) shows, these replications are no panacea for testing the veracity of psychological phenomena. If anything, they create more conceptual and methodological problems than the original studies. Baumeister et al. (35) have recently reported that these methodological problems include, but are not limited to, operational failures to test the hypothesis (as indicated by the manipulation checks on the independent variable between the original and replication studies), non-sensitive dependent variables to detect the effect, low engagement and interest among study participants, and high exclusion

rates of participants, all of which have led to weak tests of the hypothesis in the multilab replication studies.

Ego depletion has become one of the most frequently studied and "the most storied" phenomena in psychology during the last 20+ years, since its introduction by Baumeister et al. in 1998 (8). This is not surprising given that the effect is one of the most important phenomena in all of psychology, as it is at the core of human existence and functioning. Daily living constantly requires us to exercise self-control: do not smoke, do not eat unhealthy food, do not drink excessively, do not watch TV but go for a walk/run, do not use profanities in public or make sexist and racist comments etc. All of this wears people down because self-control resources can individually be finite (38), leading people to say: "enough is enough", throwing out the proverbial towel and giving up.

The ego depletion effect has almost exclusively been tested in lab experiments and has now become a frequent target of preregistered multilab replications. One of the latest is an attempt by Vohs et al.'s (9) 36-multilab study to test the effect in six countries involving 128 researchers from almost as many Universities. Despite utilizing 128 authors, this replication study has major conceptual, theoretical and methodological problems. It should be noted that their study is no different than other replication attempts using meta-analyses (e.g., 10,11,12), all of which have been marred by glaring deficiencies in theory, methodology, and statistics.

First, no theoretical rationale was given as to why the experimental tasks were selected for testing the effect. There is no theoretical justification to assume that any task A would have a self-control depleting effect on any task B. The authors from different labs were simply asked to indicate "how effective they believed the tasks would be for testing ego depletion", in other words, just their subjective opinions, but no well-developed theoretical reasons. Two different task protocols were used (i.e., two task As and two task Bs), E-task and writing-task protocols. In the latter, participants wrote a story about their recent trip for five minutes (task A) and then answered "cognitive estimation" questions (task B), such as "How many seeds are there in a watermelon?" Why would the former task have anything to do with answering silly questions about watermelon seeds, and why would a simple task of writing a story for five minutes be self-control depleting? It might be more effortful to write something for five minutes than not write anything at all, but it certainly would not be enough to deplete self-control resources. In contrast, when students write for 1-2 hours in their exams, it can easily be ego depleting.

Furthermore, it has been suggested that ego depletion should be tested in tasks that are meaningfully related to people's behaviors (13). One way to do this would be to measure self-control depletion following 8 hours of work (task A) and then determine the depletion effect on workers' subsequent ability to resist temptations in leisure (e.g., time spent watching TV vs. engaging in some form of exercise, task B) (49), especially given that the utility of leisure in part derives from the relief it provides from costly cognitive control (50). Instead, in Vohs et al.'s replication study, meaningless or random tasks were used to test the effect. Yet, the sensitivity of the dependent variable is critical for detection of differences in the effect between the original and replication studies (35), and thus for understanding replications' failures and successes.

The problem with replication research generally is that it ignores the validity of measurements and focuses on reliability, with the assumption that if a phenomenon is real, it can reliably be reproduced. This was evident in Vohs et al.'s study (9). There were no data

reported to indicate that the dependent variables were valid measures to detect ego depletion, nor was there any evidence that the chosen tasks were valid procedures for testing this phenomenon. It has been reported that behavioral measures are less accurate and valid for measuring the underlying mechanism for self-regulation (14). While reliability is important, validity is a more critical issue because "replicability does not equal validity" (15).

Second, as can be imagined, any replication study using 36 laboratories from six different countries is likely to lead to many methodological discrepancies between the sites. It is therefore not surprising that over 30% of the participants were excluded from the study (9) for various reasons, nor is it surprising that the laboratories chose different task protocols for the study participants. The study found significant support for ego depletion when the entire sample was analyzed but not when 30% of the sample was excluded, which raises serious questions about the meaningfulness of the findings.

Other methodological problems included the fact that the manipulation checks measured participants' perceptions of effort and task difficulty to show that the experimental task As were ego depleting. The manipulation check showed that one protocol was more effortful than the other. A problem, of course, is that tasks can be effort-demanding but not necessarily self-control depleting, especially when writing about something just for five minutes. Participants' motivation was also measured, but the results showed no difference in self-reported motivation between the depletion and non-depletion conditions. If a task A is ego depleting, it surely should also be motivation-depleting. Altogether, the manipulation checks indicated that this replication attempt was at best a weak test of the ego depletion effect.

Baumeister et al. (35) have recently shown that the manipulation checks, generally used in experiments to provide evidence for the effectiveness of the manipulation of an independent variable, have been a major problem in preregistered multilab replications, revealing "operational failures" in testing the effect. A true replication test of an effect requires that the manipulation checks would show a greater difference in the treatment effectiveness between the experimental and control groups in the replication than original study. These researchers concluded that "multilab studies with successful manipulation checks were more successful at replicating original results than the ones with failed checks" (35).

Third, in the light of the above and other conceptual and methodological problems, it is not surprising that results showed considerable variation in the effect sizes between the lab sites, ranging from 0.83 to -0.29 with an average of a nonsignificant effect size of 0.06 for the depletion effect. Such statistical averaging is misleading and inappropriate because it hides the methodological differences between the testing sites. Results further revealed that the E-task protocol showed a greater depletion effect than the writing-task protocol, especially on the duration variable (i.e., time before participants gave up on the task B). The Bayesian analyses showed no depletion effect, but there was "the large uncertainty associated with individual laboratories' effect sizes". It should also be noted that these analyses were based on an arbitrary "prior" of 0.30 obtained from splitting the difference between two earlier findings (effect sizes of 0.62 and 0.04).

Fourth, whether the ego depletion phenomenon is real or not has narrowly been defined in replication studies, always by statistical means, more specifically by the magnitude of the effect (effect size). Yet, many psychological effects are subtle (2,5) and small but "impressive" (16) and they accumulate to produce meaningful outcomes (17). The importance of a small effect is perhaps best

illustrated by the effect of little exercise for cardiovascular health. Research has shown that even 10-minute daily bouts of aerobic exercise can cumulatively lead to beneficial effects; of course, moderate and vigorous exercise is better. Detection of small effects, however, requires large sample sizes.

All the above means that psychological effects are often not manifested in statistically strong effect sizes (18). Besides the magnitude-of-effect, there are other equally important, if not more important, criteria to be considered (2): What about the frequency of the effect? Does the effect have to occur frequently, and how frequently, in human behaviors for it to be real? And, what about the effect's durability? How long does it have to last for it to be real? And, how many people (10, 100, 10 million?) have to experience ego depletion before it can be called real? And, what about ego depletion as a lab vs. real-life phenomenon? And, what about ego depletion as a conscious vs. nonconscious phenomenon? Do people become more ego depleted when doing consciousness-demanding vs. non-consciousness-demanding tasks? As with human behaviors generally, do people delegate the repeatedly experienced self-control depletion to the operations of the nonconscious mind? All these questions should be answered before anything can be said about the phenomenon's unimportance and unviability in explaining human behavior (2).

The above questions highlight the essence of psychological phenomena and associated empirical research. Consider feelings of anger. Most of the time, people are not angry, but it does not mean that anger as a psychological phenomenon does not exist. Researchers are called on to investigate factors that give rise to anger and those that reduce it, thereby seeking to establish the boundary conditions for the phenomenon.

Despite the enormous effort expended in completing Vohs et al.'s replication study (9), we are no closer to knowing whether ego depletion is a real phenomenon or not, albeit exercise in futility. The authors concluded that one of the following could be true: (1) there is no depletion effect, (2) the reliability of the effect is still unknown, (3) there may be a small effect. After employing 36 labs from around the world and 128 researchers, one might ask: Was it worth the expended effort to arrive at this conclusion that could have been drawn without any replication study? Rather than trying to determine whether (yes-or-no) ego depletion is a bona fide psychological phenomenon, it would have been much more useful for advancing knowledge on self-control failures if these 128 researchers had performed their own constructive-phenomenon replications to investigate different manifestations of self-control depletion and to try to establish boundary conditions for ego depletion. Phenomenon replications test effects in varied forms, ways, contexts and times using different methods and multiple criteria, not just the statistical magnitude of an effect (5).

It may be argued that one unintended consequence of the ego depletion replication research is that it takes investigators' attention away from a more important task, namely, to develop a broader theory of what factors account for self-control failures and successes. Ego depletion is just one of the many potential determinants of self-control failures, and it may not even be the most important of them. Although new but isolated hypotheses about the self-control depletion mechanism have been proposed (39), no theory has been advanced to explicate the self-control process and associated factors underlying self-control failures and successes, and to show how ego depletion fits with a broader theory. When emphasis is placed on proving the phenomenon's existence (yes-or-no), it discourages researchers' attempts to conduct constructive replications and undermines their creative theoretical work. As a result, an inordinate amount and number of resources are

spent trying to determine whether this one aspect of the self-control process is real or not.

Replication In Medical and Economic Sciences

Medical Sciences. The replication problem is not just a problem of psychological sciences. It is equally challenging and intractable in medical and economic sciences. In physics, scientists seek to discover the laws of *nature*, whereas in psychological and medical sciences, both *nature* and *nurture* must be considered. This interaction effect of person and environment in human condition, performance and behavior makes direct and exact replications impossible. For example, in clinical studies, a drug's efficacy is often evaluated in terms of the extended length of life in months, but seldom in years. While a new drug or therapy may be significantly more efficacious in an experimental than a placebo group, there are often considerable individual differences in these effects due to genetics and lifestyle factors. For example, the drug may not be as efficacious for smokers, non-exercisers, alcohol users, and those who follow unhealthy diets.

Rather than performing direct replications, it would be more useful to conduct constructive replication studies to determine how the drug's efficacy varies as a function of group and individual differences. Furthermore, it would be important to ascertain how drugs' effects wane with time, meaning that new studies are needed to establish the efficacy of various combinations ("cocktails") of old and new drugs and therapies. This is equivalent to the effects of anxiety on human performance. Although anxiety generally has a negative effect on human performance, its effect, however, wanes with experiences. For example, with competitive experiences, athletes learn to deal with anxiety and therefore become more proficient in mitigating its negative effects.

An interesting case is the vaccine-vaccination paradox. On one hand, with the unprecedented speed, the development of the Covid-19 vaccines led to one of the greatest achievements in the history of science, especially in medical and biological sciences. This achievement, however, was preceded by painstaking basic research over many decades, and by multiple failed experiments until the mRNA technology was seen feasible for the development of effective vaccines. What makes this case interesting is that medical and psychological sciences meet at the vaccine-vaccination intersection. Even though highly effective vaccines are now available, the human mind has interfered in the process as millions of people refuse to get vaccinated for various, questionable reasons. While human thinking can be predictable and stable, it can also be fickle, elusive and easily persuaded at times, again demonstrating the "hard" nature of psychological science. It is one and an admirable thing to develop effective therapies and treatments to improve the human condition, but it is altogether a different matter to get people to use them.

It is noteworthy that Francis Collins, NIH director, is now lamenting the fact that people are not rational in their refusal to get vaccinated. He supported medical and biological research to get vaccines developed but cut funding for psychological and behavioral research that would have shed light on why people are often irrational in their thinking and behaviors. If we are unable to persuade most people to get vaccinated even when facing serious consequences from not doing it, what hope is there for getting the 78% segment of the population that is sedentary to start exercising regularly? (19,20,21). A lot of original studies and constructive replications remain to be funded and conducted.

Economic Sciences. The situation is no better in economic sciences. Human conditions vary, for one thing, because individuals (investors, policymakers, and politicians) are

not rational and invariant in their decisions and judgments. "Behavioral economists", led by two Nobel Prize winners (Kahneman and Thaler), have shown that the rational-agent model is a poor explanation for financial decisions, asset prizes, and economic growth more generally. Investors may collectively consider all the "known" information in decision-making, as assumed by the prevalent "Efficient Market Hypothesis", but individually, they are influenced by self-generated and situationally engendered emotions when making financial decisions. Furthermore, individual investors (e.g., prospective home buyers) can be led to make irrational decisions, resulting in "collective blindness" (22) that in turn can cause national and international financial crises, as seen in the 2008 financial calamity.

All the above means that there will be deviations from the general patterns of individual financial behaviors and their underlying conditions, making direct and precise replications not feasible. However, conceptual or phenomenon replications are helpful in elucidating the boundary conditions of overall patterns (5,40), that is, conditions under which phenomena are strong vs. subtle, stable vs. variable. But if we insist on precise replications, then no psychological and economic phenomena exist because it is impossible to reproduce conditions identical to those of the original testing.

Consider the famous Phillips Curve and its recent failure to explain the negative relationship between unemployment and inflation. Even though the Curve has served as the basic tenet of economics for over 60 years and guided the Federal Reserve in its policy decisions, many economists have recently rejected the phenomenon as inflation has been close to zero during the last 10 years or so. But now that inflation has surged and unemployment declined dramatically, the Phillips Curve may again be invoked to explain economics at the macro level. Had economists done conceptual replications policymakers would have been better informed about the power and limits of the Phillips Curve as an explanatory mechanism. Although inflation as an economic phenomenon has not existed in recent years, it would be foolhardy to try to prove the nonexistence of the Phillips Curve in replication studies. Instead, studies should be done to determine what makes this phenomenon fluctuate, thereby seeking to establish its boundary conditions.

In sum, there are no universal constants to be precisely replicated outside of the laws of nature and physics. If the conditions are not the same at the individual level from one experimental situation to another, they are not the same at the macro level either. Besides, history does not exactly repeat itself, it only rhymes, as Mark Twain so eloquently put it. The conditions that led to a recession at one time will not be the same causes for the next recession. At the macro level, researchers may build theoretical models trying to predict the next recession, but they cannot conceivably consider all the relevant variables, especially exogenous ones, and thus precise predictions (and exact replications) are not possible. Regardless, this has not prevented pundits from declaring that it is "different this time" or that it is "a new normal".

Conclusion

Theoretical and methodological deficiencies cannot be saved by fancy statistical analyses or the pinpoint statistical verification of an effect through replication. Science mainly advances by theory building and model construction and expansion, not by repeated empirical tests and replications of the statistical null hypothesis. Psychological, medical and economic phenomena are largely theoretical constructs, not unlike those in physics. Just think where physics and the world would be today without Einstein's theories (e.g., no GPS). In particle physics, the Higgs boson particle was theorized to exist in

1964 but not verified until 2012. Did the particle not exist in the meantime?

Empirical studies are mainly evaluated for their theoretical relevance and importance, not their success or failure to exactly reproduce the original findings. It is not the empirical data but theory that has generally made scientific progress possible, which is as true of physics as of psychological, medical and economic sciences. Along the way, empirical data have complemented and contributed to the clarification and expansion of theoretical models, and theories have made data more useful and informative, in concert with Einstein's famous maxim (extended from religion to science): "Data without theory is lame; theory without data is blind". Of course, extant theories are eventually abandoned in the Kuhnian-like paradigm shifts. But in the meantime, researchers pursue "temporary winners" as scientific knowledge is preliminary, conditional, provisional and propositional.

Ego depletion is alive and well. No failed replication can ever wipe it off the face of earth. There will always be people who experience it in one form or another, and it will affect them in differing ways and to differing degrees, either emotionally, cognitively or behaviorally. Moreover, by 2019, there were about 600 empirical studies reporting support for the ego depletion effect, including some preregistered multilab studies (e.g., 41,42), making ego depletion one of the four most replicated phenomena in social psychology (35); the other three are mortality salience, Elaboration Likelihood Model of persuasion, and priming, with each having over 100 successful replications (35). What is disturbing from the standpoint of reporting scientific findings is that journal editors have shown blatant bias toward favoring nonreplicated findings, with the editor in the case of the Vohs et al. study directing the authors to report the nonsignificant results (failure to replicate) in the published article but directing them to place the significant results (success to replicate) in the Supplementary Online materials (35). Relatedly, there are numerous examples in the literature of methodological liberties replication researchers have taken in attempts to disprove the original findings. These "replicator degrees of freedom" have been shown to lead to unwarranted claims of replication failures (43,44,45).

The replication pendulum has swung to a destructive extreme mostly because of the emphasis on the dichotomous statistical decisions regarding effects' existence and the statistical determination of effects' strength and magnitude while ignoring other criteria. Scientific inferences go far beyond statistical inferences. It is time to swing the pendulum back to the middle by conducting constructive-phenomenon replications that test phenomena in varied forms, contexts, and times using different methods. Such replications provide more nuanced and refined explanations than categorical declarations that phenomena are or are not real. These replications use multiple criteria other than the statistical magnitude of the effect and shed light on phenomena's boundary conditions in the ongoing pursuit of temporary scientific truth.

Acknowledgement

The author thanks John Bargh, Roy Baumeister, Bradley Hatfield, and Matthew Miller for their most helpful and constructive comments and suggestions on an earlier version.

References

1. Gleiser M. The island of knowledge. 2014. New York, NY: Basic Books.
2. Iso-Ahola S. Reproducibility in psychological science: When do psychological phenomena exist? Frontiers in Psychology. 2017; 8 (879): 1-16.
3. Anderson C, Bahnik S, Barnett-Cowan M, Bosco F, Chandler J, Chartier C, et al. Response to comment on "Estimating the reproducibility of psychological science". Science. 2016; 351: 1037.
4. Earp B, Trafimow D. Replication, falsification, and the crisis of confidence in social psychology. Frontiers in Psychology. 2015; 6: 621.
5. Iso-Ahola S. Replication and the establishment of scientific truth. Frontiers in Psychology. 2020; 11 (2183): 1-15.
6. Rubin M. What type of Type I error? Contrasting the Neyman-Pearson and Fisherian approaches in the context of exact and direct replications. Synthese. 2019; 196: 1-26.
7. Stroebe W, Strack, F. The alleged crisis and the illusion of exact replication. Perspectives in Psychological Science. 2014; 9: 59-71.
8. Baumeister R, Bratslavsky E, Muraven M, Tice D. Ego depletion: Is the active self a limited resource? Journal of Personality and Social Psychology. 1998; 74: 1252-1265.
9. Vohs K, Schmeichel B, Lohmann S, Gronau Q, Finley A, Ainsworth S. et al. A multiple preregistered paradigmatic test of the ego-depletion effect. Psychological Science. Psychological Science. 2021; 32: 1566-1581.
10. Carter E, Kofler L, Forster D, McCullough M. A series of meta-analytic tests of the depletion effect: Self-control does not seem to rely on a limited resource. Journal of Experimental Psychology: General. 2015; 144: 796-815.
11. Hagger M, Chatzisarantis N, Zwienenberg M. A multilab preregistered replication of the ego depletion effect. Perspectives on Psychological Science. 2016; 11: 546-573.
12. Hagger M, Wood C, Stiff C, Chatzisarantis N. Ego depletion and the strength model of self-control: A meta-analysis. Psychological Bulletin. 2010; 136: 495-525.
13. Iso-Ahola S. Conscious versus nonconscious mind and leisure. Leisure Sciences. 2015; 37: 289-310.
14. Enkavi A, Eisenberg I, Bissett P, Mazza G, MacKinnon D, Marsch L, et al. Large-scale analysis of test-retest reliabilities of self-regulation measures. Proceedings of National Academy of Sciences USA. 2019; 116: 5472-5477.
15. Hussey I, Hughes S. Hidden invalidity among 15 commonly used measures in social personality psychology. Advanced in Methods and Practice in Psychological Science. 2020; 3: 166-184.
16. Prentice D, Miller D. When small effects are impressive. Psychological Bulletin. 1992; 112: 160-164.
17. Funder D, Ozer D. Evaluating effect size in psychological research: Sense and nonsense. Advances in Methods and Practices in Psychological Science. 2019; 2: 156-168.
18. Meehl P. Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant it. Psychological Inquiry. 1990; 1: 108-141.
19. Iso-Ahola S. Exercise: Why it is a challenge for both the nonconscious and conscious mind. Review of General Psychology. 2013; 17: 93-110.
20. Iso-Ahola S. Conscious-nonconscious processing explains why some people exercise but most don't. Journal of Nature and Science. 2017; 3 (e384): 1-16.

21. Iso-Ahola S. Human mind: Both the cause and solution to the global pandemic of physical inactivity. International Journal of Public Health Research. 2018; 6: 107-113.
22. Kahneman D. Thinking, fast and slow. 2014. New York, NY: Farrar, Straus and Giroux.
23. Trafimow D. Hypothesis testing and theory evaluation at the boundaries. Psychological Review. 2003; 110: 526-535.
24. McFall R. Making psychology incorruptible. Applied and Preventive Psychology. 1996; 5: 9-15.
25. Popper K. The logic of scientific discovery. 1959. London: Hutchison.
26. Lykken D. Statistical significance in psychological research. Psychological Bulletin. 1968; 70: 151-159.
27. Ackerman P. Nonsense, common sense, and science of expert performance: Talent and individual differences. Intelligence. 2014; 45: 6-17.
28. Bargh J. The cognitive unconscious in everyday life. 2021. In A Reber, R Allen (Eds), The cognitive unconscious. London: Oxford University Press.
29. Cumming G. The new statistics: Why and how. Psychological Science. 2014; 25: 7-29.
30. Cohen J. Statistical power analysis for the behavioral sciences. 1988. Hillsdale, NJ: Erlbaum.
31. Wetzels R, Matzke D, Lee M, Rouder J, Iverson G, Wagenmakers E-J. Statistical evidence in experimental psychology: An empirical comparison using 855 t tests. Perspectives on Psychological Science. 2011; 6: 291-298.
32. Kass R, Raftery A. Bayes factors. Journal of American Statistical Association. 1995; 90: 377-395.
33. Dienes Z. Bayesian versus orthodox statistics: Which side are you on? Perspectives on Psychological Science. 2011; 6: 274-290.
34. Bargh J. The historical origins of priming as the preparation of behavioral responses: Unconscious carry-over and contextual influences of real-world importance. In D Molden (Ed), Understanding priming effects in social psychology (pp. 218-233). New York, NY: Guilford Press.
35. Baumeister R, Tice D, Bushman B. A review of multi-site replication projects in social psychology: Methodological ideal or collective self-destruct mechanism? 2022. Submitted for publication.
36. Churchland M, Afshar A, Shenoy K. A central source of movement variability. Neuron.2006; 52: 1085-1096.
37. Bargh J, Hassin R. Human unconscious processes in situ: The kind of awareness that really matters. 2021. In A Reber, R Allen (Eds), The cognitive unconscious. London: Oxford University Press.
38. Baumeister R. Conquer yourself, conquer the world. Scientific American. 2015; 312: 61-65.
39. Inzlicht M, Schmeichel B. What is ego depletion? Toward a mechanistic revision of the resource model of self-control. Perspectives on Psychological Science. 2012; 7: 450-463.
40. Crandall C, Sherman J. On the scientific superiority of conceptual replications for scientific progress. Journal of Experimental Social Psychology. 2015; 66: 93-99.
41. Dang J, Barker P, Baumert A, Bentvelzen M, Berkman E, Buchholz N, et al. A multilab replication of the ego depletion effect. Social Psychological and Personality Science. 2021; 12: 14-24.
42. Garrison K, Finley A, Schmeichel B. Ego depletion reduces attention control: Evidence from two high-powered preregistered experiments. Personality and Social Psychology Bulletin. 2019; 45: 728-739.
43. Bryan C, Walton G, Rogers T, Dweck C. Motivating voter turnout by invoking the

self. Proceedings of National Academy of Sciences USA. 2011; 108: 12653-12656.
44. Bryan C, Yeager D, O'Brien J. Replicator degrees of freedom allow publication of misleading "failures to replicate". Unpublished manuscript. 2019. University of Chicago.
45. Ramscar M, Shaoul C, Baayen R. Why many priming effects don't (and won't) replicate: A quantitative analysis. Unpublished manuscript. 2015. Tubingen University, Germany.
46. Chen X, Latham G, Piccolo R, Itzchakov G. An enumerative review and a meta-analysis of primed goal effects on organizational behavior. Applied Psychology. 2021; 70: 216-253.
47. Weingarten E, Chen Q, McAdams M, Yi J, Hepler J, Albarracin D. From primed concepts to action: A meta-analysis of the behavioral effects of incidentally presented words. Psychological Bulletin. 2016; 142: 472-497.
48. Bargh J. The hidden life of the consumer mind. Consumer Psychology Review. 2021; 5: 1-16.
49. Hofmann W, Vohs K, Baumeister R. What people desire, feel conflicted about, and try to resist in everyday life. Psychological Science. 2012; 23: 582-588.
50. Kool W, Botvinick M. A labor/leisure tradeoff in cognitive control. Journal of Experimental Psychology: General. 2014; 143: 131-141.
51. Amrhein V, Greenland S, McShane B. Retire statistical significance. Nature. 2019; 567: 305-307.
52. Greenland S, Senn S, Rothman K, Carlin J, Poole C, Goodman S. et al. Statistical tests, p values, confidence intervals, and power: A guide to misinterpretations. European Journal of Epidemiology. 2016; 31: 337-350.
53. Iso-Ahola S, Dotson C. Psychological momentum--not a statistical but psychological phenomenon. Review of General Psychology. 2015; 19: 112-116.

