Re-Inquiries

The Desperate Need for Replications

JOHN E. HUNTER*

An overemphasis on creativity for evaluating research has lead to a serious devaluation of replication studies. However, we need a total sample size of N = 153,669 to estimate a causal effect to two digits, which is quite rare for a single study. The only way to get accurate estimation is to average across replications. If the average sample size were as high as N = 200, we would need over 700 replication studies. Scientific replications are more problematic than pure statistical replications, and so we need even more replications to achieve reasonable accuracy.

I have been very disturbed by articles rejected as mere replications of existing studies. In this essay, I intend to point out the grave errors in that judgment. The fact is that we need to publish replication studies of all types and we need many such replications for each study.

The basic error in rejecting replications as redundant stems from the combination of two other errors. First, there is no problem in using creativity as a criterion for quality of scientific studies. In part, scientific progress depends on new ideas, and a stress on creativity generates the search for new ideas. However, scientific progress also requires a database of facts, and facts can only be established by replicated studies. The error in rejecting replications is using creativity as the only criterion for publication and ignoring the importance of facts.

The second error in rejecting replications is the widespread belief that single studies establish findings and, thus, a replication adds nothing. Statistical reasoning alone will show this belief to be very wrong. This weakness in single study results has been widely recognized implicitly by the widespread acceptance of meta-analysis. There is a definite trend toward using meta-analysis as the best form of fact-finding literature review. Textbook authors increasingly use meta-analysis as their prime citation.

There are three kinds of replication studies, and all three kinds are desperately needed. The three kinds of studies can be listed in order of increasing creativity: (*a*) statistical replications, (*b*) scientific replications, and (c) conceptual replications. I will argue the necessity of replications in that order, starting with the least creative of replications.

*John E. Hunter is professor of psychology at Michigan State University, East Lansing, MI 48824 (hunterj@msu.edu). His research interests include personnel selection, organizational behavior, and meta-analysis.

STATISTICAL REPLICATIONS

Consider simple causal studies that seek to answer the question, "How large is the effect of X on Y?" where X is some specified independent variable and Y is some specified dependent variable. The simplest outcome statistic for such studies is the correlation coefficient. I will state the argument for replications for the correlation because it is the simplest outcome statistic. However, any statistician will tell you that the argument holds for all sample statistics. In a nutshell, we need statistical replications in order to reduce sampling error to a small enough level to draw correct conclusions. For statistical replications as perfectly replicated studies:

a) All studies measure the independent variable in exactly the same way.

b) All studies measure the dependent variable in exactly the same way.

c) All studies use exactly the same procedure.

d) All studies draw samples from the same population.

To a person wedded to creativity as the only criterion for scientific quality, such studies would appear to be completely redundant because they are substantively identical. How could a second study add any meaningful knowledge to the first study done?

One major problem with any single study is sampling error. Most researchers report a correlation rounded to two digits. If this is the required accuracy for knowledge of a

correlation, we can compute the sample size needed so that the sample correlation will come out correctly to two digits. That sample size is a staggering $N = 153{,}669$. For the correlation to be accurate to within one digit, the sample size must be $N = 1{,}544$. Only a rare social science study has such a sample size. Furthermore, even if a large survey study had a total sample size of $N = 1{,}544$, there would be many subset analyses where the sample size would be much smaller.

The *Journal of Consumer Research* has an average sample size of about $N = 200$. For one-digit accuracy, these studies should be more than seven times larger. Since the size of social science studies is usually determined by maximal constraints on gathering data, most researchers cannot increase the size of their studies. Indeed Sedlmeier and Gigerenzer (1989) found that the average sample size in psychology had not increased in 30 years.

If we want one-digit accuracy in estimating basic facts in the social sciences, then we must have more than one study. If the average sample size is $N = 200$ for single studies, then we need eight statistical replications in order to obtain the needed total sample size. If we want two-digit accuracy, we need 800 replications.

Every person who is familiar with the facts of sampling error has urged scientists to do and report replications. The replications they have in mind are the perfect statistical replications of statistical theory, studies that are devoid of substantive creativity after the first one is done. All scientists should be aware of these facts.

SCIENTIFIC REPLICATIONS

Scientific replication studies are rarely pure statistical replications. Indeed a statistician will often fail to recognize such studies as replications. For scientists, replicated studies are studies that are scientifically equivalent to each other. For simple causal studies, we can lay out the requirements for scientific replications. For scientific replications for simple causal studies:

a) All studies measure the same independent variable *X*.

b) All studies measure the same dependent variable *Y*.

c) All studies use essentially the same procedure.

d) All studies should sample from populations that are equivalent in terms of the study question and hence the study outcome.

The first three assumptions for scientific replications are very similar to the first three assumptions for perfect replications. The difference is that statistical replications assume that the word "same" means identical, while scientists interpret the word "same" to mean equivalent.

Consider the independent variable. Scientific replications should all measure the same independent variable. Perfect replications guarantee that by assuming that the measurement procedures are identical across studies. Scientific replications allow equivalent measurements across studies. Even though the specific measures of the independent variable are different, it is assumed that the various measures are construct equivalent. That is, free of random error of measurement, all measures of the independent variable would be perfectly correlated with each other. In many areas, this assumption can be questioned by critics of the results.

Consider the dependent variable. The same distinction applies: perfect replications guarantee equivalence by requiring identical measurement methods across studies, while scientists require only equivalent measurement across studies, that is, measures that might be different in detail but would show perfect congruent validity with each other. A critic might argue this assumption.

Consider procedures. Scientific replications must use procedures that are functionally the same for different studies. Perfect replications guarantee this by assuming identical procedures across studies. Scientists would not require that studies be identical on irrelevant details. A critic might argue that certain irrelevant details are actually critical to the outcome for the study.

The assumption about populations is different from the other three assumptions. In the context of simple causal studies, most scientists have not thought very much about differences between subject populations. This lack of attention in simple causal studies stems from the current weakness in theories concerning individual differences. Few theories address the issue of whether the causal effect would be different for different kinds of people.

The one area that has received considerable attention is the difference between humans and other species, and there is now a very considerable literature on the differences between animal species. For example, many findings in the medical literature do not generalize from mice to rats, much less from mice to humans. In drug research, a positive finding for mice is treated as a suggestion that the finding might generalize to humans, but no one seriously considers the findings on mice as definitive.

It is important to keep these medical findings in perspective. The fact that certain findings do not generalize from mice to humans shows that you must test the findings on mice with a study on humans. However, these findings that fail to generalize are famous precisely because they are so peculiar. Most chemical findings do generalize from mice to humans. That is why most drug studies start with mice.

Consider simple causal studies in the social sciences. The key to generalization from one population to another has to do with the causal effect being studied. If that effect is the same in the two populations, then the two populations are equivalent even though the populations might be different on other irrelevant dimensions.

Meta-analysis was developed to analyze and compare results across replicated studies. Most such domains are scientific replications rather than statistical replications. Some critics have vehemently condemned meta-analysis in such domains. Their argument is now called the apples-and-oranges argument against meta-analysis. Apples-and-oranges critics of meta-analysis claim that if studies differ in any way, then those studies are not replications, and it makes no sense to estimate results by combining across studies that are not comparable to one another. That is, they implicitly

claim that meta-analysis can only be done with perfect replications.

It is important to note that the critics do not start from data. For example, no critic yet has presented evidence that different dependent variable measures in a given domain actually measure different variables. Rather they just assume that since the measures differ, they cannot be measures of the same construct. This assumption has been proven false for many important constructs where different measures have been empirically shown to be construct equivalent.

A domain of scientific replications has a further problem beyond a domain of statistical replications. Critics can argue that the studies are not replications of each other and thus should not be combined in order to reduce sampling error. There are empirical solutions to this problem, but those solutions require more data than are required by statistical replications.

We will first consider the critic's argument in the context of perfect studies. In that context, we will show that the critic's argument can fall completely apart. It is possible for studies that do not look like perfect replications actually to be perfect replications anyway.

We will then consider both sides of the argument in the context of imperfect studies. The argument is much more complicated for both sides when we acknowledge the fact that there are no perfect studies.

Perfect Studies

While scientists pride themselves on being realistic, there are many surprisingly romantic ideas that flourish in science. The myth that creates the most problems for discussions of methodology is the myth that a scientist can do a perfect study. This myth stems from feelings of pride in hard work that is well done. Surely if a scientist is intellectually, personally, and morally committed to state of the art research methods, then that scientist can do a perfect study.

The myth of the perfect study can be destroyed by considering sampling error via Monte Carlo simulations. Even if studies are substantively perfect, there will usually be some random element in the research design—such as the choice of subjects who participate in the study. If there is a random element in the study design, then the study outcome will have sampling error that will only vanish for studies of enormous size.

For this discussion, we assume substantively perfect studies but acknowledge the effect of sampling error on study outcomes. Because we admit to sampling error, we know that if we could find perfect replications, then we could combine results across studies to reduce the size of the sampling error in our combined results. The question is this: When can we combine results?

If a substantive study could be done perfectly, then the independent variable would be measured perfectly. In this case, all studies seek to measure the same independent variable, and if the studies are perfect, then they will measure exactly the same variable however different the measurement procedures may appear.

Consider an analogy with measuring length. One ruler may be red while the other ruler is green, but the distance measured is the same, and difference in ruler color is irrelevant to the study outcome. If all studies have perfect measures of length, then all studies will be functionally identical even though the superficial measuring instruments seem different.

The same argument applies to the dependent variable. If the studies are substantively perfect, then they will all measure exactly the same variable as the dependent variable.

In order for a study to have a perfect procedure, the researcher must correctly know which procedural details are critical and which are incidental. All perfect studies will use only the critical elements of the procedure and will eliminate any aspect that might alter the fundamental nature of the procedure. That is, if all studies in a domain are substantively perfect, the procedures will all do exactly the same thing.

The same argument would apply to the study population. If researchers knew that the treatment effect is the same in a wide variety of populations, they could draw a sample from any such population and be assured that the population effect size would be the same in all samples. This argument is on poorer grounds than other arguments because current theories are usually weak on the topic of individual differences, and thus current researchers rarely actually think about this issue.

It should be noted that there is now a massive body of empirical evidence on this issue. Almost all major findings about causal effects have been found to generalize across human populations. Indeed most results generalize across species of mammals.

Consider simple causal studies in the social sciences. The key to generalization from one population to another has to do with the causal effect being studied. If that effect is the same in the two populations, then the two populations are equivalent even though the populations might be different on other irrelevant dimensions.

This has been the finding in most meta-analyses. I know of no meta-analysis that has found study location to be a major moderator variable. This issue has been studied in detail in industrial psychology in connection with the validity of cognitive and personality measures predicting job performance. There is evidence that results differ from one job to another, but no evidence for differences in results for the same job across different locations.

The one exception is studies on specially selected populations. If one study looks at personality in a full normal population, while a second study considers only psychiatric patients, the results will often not be comparable. These issues of biased sampling have been studied in psychometric theory for 80 years now, and there are methods for correcting results for differences in sample selection. These methods require researchers to pay great attention to standard deviations and are thus unknown to most current researchers who ignore standard deviations. This is a big problem for meta-analysis since current researchers do not even

publish standard deviations. It has been my experience that when you ask for a much larger technical report, the standard deviations are not there either.

The upshot of this argument is this: if studies were substantively perfect, they would be statistical replications even though they do not look like statistical replications. The sampling error arguments presented above apply exactly to substantively perfect studies. The need for replications in perfect studies is precisely to increase sampling error; there is no other problem. However, it is important to know that if the average sample size in a single study is $N = 200$, then we need 800 replications for perfect studies to generate a result that is accurate to two digits.

Almost Perfect Studies

If we have substantively imperfect studies (and I believe the empirical evidence has proven that all current studies are imperfect), then the apples-and-oranges criticism of replicated studies cannot be answered by a purely substantive argument. The question is this: How can we look at study results to see if the studies are replications or if they actually differ from one another on some unexpected critical dimension?

The simplest case assumes that studies are substantively perfect except for the procedure. The apples studies use a treatment that incorporates a critical detail, while the oranges studies miss that detail. The effect size should differ substantially between the two kinds of studies. That hypothesis can be tested empirically.

Consider the beliefs of the reviewer who gathers the studies. The reviewer believes that most researchers in the area know which procedural details are critical and that all of the studies selected for the review are correct in those details. Thus the reviewer believes that all studies will have the same outcome and will thus function as statistical replications.

By contrast, the apples-and-oranges critic thinks that researchers are mostly ignorant and have little idea what details are critical and what details are incidental. Thus the critic expects study outcomes to differ randomly from one study to the next. In particular, the critic expects the population effect sizes to vary sharply across studies.

This issue has received considerable study in meta-analysis and has led to widespread use of the terms "homogeneous" versus "heterogeneous" for a study domain:

DEFINITION: A domain is called "homogeneous" if the population study effect sizes are uniform across studies.

DEFINITION: If population values vary across studies, the domain is said to be "heterogeneous."

WARNING: The definitions above are not as simple as they appear. These definitions are usually made by people who believe in perfect studies. For imperfect studies, there is a new complication that will be discussed below.

For a set of studies with perfect measurement and equivalent study populations, the reviewer's hypothesis of equivalent procedures leads to the hypothesis that the results will be homogeneous on that domain. By contrast, the critic predicts that the domain will be wildly heterogeneous.

Meta-analysis can test a series of study results to see if the population values are uniform across the domain. The poorest way to test the hypothesis is with a significance test called the "homogeneity test" by some and the "heterogeneity test" by others. If the test registers not significant, then the domain is labeled homogeneous. If the test registers significant, then the domain is labeled heterogeneous.

Assume that the significance test works. Then from the data at hand, we can test the hypothesis that the studies are actually statistical replications. If the studies are replications, then the domain will be homogeneous and the significance test will say not significant. This will confirm the hypothesis generated by the reviewer in reading and selecting the studies.

By contrast, suppose the reviewer has made an error. Certain studies that the reviewer thought to be equivalent to each other are actually different from each other. The domain would be heterogeneous, and the significance test result would be significant. The reviewer must then go back to the studies and look for an error in the review analysis.

The problem with the significance test is the problem with all significance tests. If the null hypothesis is true, then the significance test works with the advertised low error rate (usually 5 percent). That is, if the only problem is type I error, then the significance test has only a 5 percent error rate.

The big problems with significance tests come if you must worry about type II error. If the domain is heterogeneous, then a type I error is impossible; you cannot possibly falsely claim that a domain is heterogeneous if it is in fact heterogeneous. For a heterogeneous domain, the significance test is wrong if it says not significant. This would falsely suggest that the domain is homogeneous when it is actually heterogeneous. Since the heterogeneity test is a one-way test, there is a potential error rate of 95 percent for type II error.

There is only one solution to the power problem for this significance test: we must have many studies (even if the individual studies have a large average sample size). It takes even more replications to show homogeneity than to estimate the population effect size!

Real Imperfect Studies

If studies are imperfect, it is likely that they will be imperfect to different degrees on different dimensions. That is, even though the studies are scientific replications, the studies will differ in quality and will hence differ in outcome.

If all studies were substantively perfect, the explanation of heterogeneity would be straightforward. If study results differ, then there must be substantive differences between the studies, and the effect size differs for different kinds of studies.

We now know that there are no substantively perfect stud-

ies. Not only are studies imperfect because of sampling error, they are imperfect because measurement is imperfect, in some areas studies may suffer from biased sampling, in some areas variables are artificially dichotomized, and so on. Psychometricians have been looking at study imperfections for nearly 100 years, and many such dimensions have been considered. John Hunter and Frank Schmidt (1990) have identified 13 dimensions on which studies are known to be potentially imperfect.

All dimensions for imperfection are known to be quantitative. Studies can vary from perfect to nearly perfect to OK to nearly awful to awful. Since imperfection is quantitative, it can be measured. We can talk about the quality of a study and can measure that quality. Study results are very much affected by study imperfections. Imperfections distort the data, and psychometricians have studied those distortion processes. The lower the quality of the study, the more distorted the study results. As a general rule, distortion has the effect of reducing the size of the observed treatment effect; the lower the quality of the study, the smaller the observed treatment effect.

If studies differ in study quality, then the study effect sizes will differ. Other things being equal, the higher quality studies will have larger effect sizes. This means that if studies differ in quality, then the population effect sizes will differ correspondingly. That is, if studies differ in quality, then the domain will be heterogeneous by the usual definition regardless of whether the actual effect sizes differ.

This can be easily illustrated with studies that differ in reliability. Suppose one study has a lot of measurement time and uses long scales for both the independent and the dependent variable. A second study has severely limited measurement time and so uses short scales to measure both variables. The effect sizes will differ because of differences in the extent of attenuation produced by the different amounts of random error of measurement in the two studies.

EXAMPLE:

Perfect measurement: effect size = $\rho = .50$.
Imperfect measurement: effect size = ρ_o.

$$\text{Long scales}: r_{xx} = r_{yy} = .81$$
$$\rho_o = \sqrt{r_{xx}}\sqrt{r_{yy}}\rho = (.90)(.90)\rho$$
$$= (.81)(.50) = .405.$$

$$\text{Short scales}: r_{xx} = r_{yy} = .64$$
$$\rho_o = \sqrt{r_{xx}}\sqrt{r_{yy}}\rho = (.80)(.80)\rho$$
$$= (.64)(.50) = .320.$$

In both cases, the imperfect study underestimates the effect size. However, the higher quality study generates a population study value of .40, while the lower quality study generates a population study value of only .32. This is typical of all known dimensions for study imperfection (i.e., not just measurement imperfections): the higher quality studies will have larger study effect sizes than the lower quality studies.

Since studies that differ in quality will differ in the size of the study population result, even a set of substantively perfect replications will not generate a homogeneous domain. Rather, the study results will differ because of differences in study quality. This means that it is very important to distinguish between the conceptually homogeneous domain of scientific replications versus the visibly homogeneous domain of perfect studies.

Warning about Definitions

When a scientist asks, "What is the effect of stress on anxiety?" the conceptual answer is given by a perfect study. If the actual effect size is the same in all study contexts, then that domain is conceptually homogeneous. However, if the studies differ in quality, then the high quality studies will have larger study population correlations than will the low quality studies. The domain will then be called heterogeneous by the usual definition current in meta-analysis.

If I could have my way, I would put the conceptual definition first and invent new terminology for bare-bones meta-analysis. However, the current definition for homogeneity is fixed in cement, and I will use that terminology. What I offer is the phrase "conceptual homogeneity" for the theoretically meaningful definition.

DEFINITION: A domain is conceptually homogeneous if the effect sizes would be uniform for substantively perfect studies.

There is no known domain with substantively perfect studies, and I do not think there will ever be such a domain in the social sciences. The question is this: How would conceptual homogeneity be affected by study imperfections?

Consider first the case of perfect replications. Each study outcome is distorted by imperfections in the study design. However, because each study has exactly the same imperfections, the distortion will be the same for each study. Therefore, all study outcomes are distorted to the same reduced value. The study values are thus still uniform across studies.

KEY FACT: If a domain is conceptually homogeneous and the studies are perfect replications, then the domain will be homogeneous by the usual definition.

By contrast, suppose we only have scientific replications. The various independent variables all measure stress, but they measure at different levels of quality (such as differences in reliability of measurement). The study treatment

correlation will then be higher for studies with higher quality measures of the independent and dependent variables. Since the study values differ, the domain is not homogeneous, but heterogeneous.

KEY FACT: If a domain is conceptually homogeneous but the studies differ in quality, then the population study effect sizes will differ across studies, and the domain will be heterogeneous.

This is a quantitative distinction. If two studies differ only trivially in quality, then they will also differ only trivially in outcome value. In many domains, it is likely that studies will be approximately uniform in quality. In that case, we will have an approximate translation from conceptual homogeneity to technical homogeneity using bare-bones meta-analysis.

KEY FACT: If a domain is conceptually homogeneous and studies are approximately uniform in quality, then the domain of study outcome values will be approximately homogeneous.

KEY FACT: If a domain is conceptually homogeneous but studies differ sharply in quality, then there will be considerable departure from homogeneity in population study outcome values.

Full Meta-analysis

Consider the distinction between full meta-analysis and bare-bones meta-analysis. Bare-bones meta-analysis makes no attempt to correct study imperfections other than sampling error. In particular, no correction is made to remove the distortion produced by measurement error. Full meta-analysis corrects as many artifacts as are relevant (assuming data are collected to measure study quality). Differences in study quality are eliminated by the full meta-analysis corrections. Thus, a domain that is labeled homogeneous by full meta-analysis is a domain that would be homogeneous if all studies were perfect studies. That is, in full meta-analysis, the word "homogeneous" means the same as what we have here called "conceptually homogeneous."

Bare-bones meta-analysis makes no corrections for study imperfections, and thus any differences in study quality result in differences in study outcome. Bare-bones meta-analysis would not find a domain to be homogeneous unless the studies were uniform in quality.

The impact of study imperfections is quantitative. If studies are approximately equal in quality, then all outcomes are distorted to approximately the same extent. Bare-bones meta-analysis would then find only very small departures from homogeneity. Furthermore, if the researcher uses the significance test for homogeneity, it is very likely that the significance test will fail to detect the departure and will register the domain as homogeneous. This is a case where the significance test would make an error at the level of bare-bones meta-analysis but might be perfectly correct at the conceptual level.

Full Meta-analysis and Data Requirements

To answer fully the apples-and-oranges critic, the reviewer must do a full meta-analysis to control for differences in study quality. A bare-bones meta-analysis will confuse differences in study outcome (because of differences in study quality) with differences in real outcome. However, if a full meta-analysis finds the domain to be homogeneous, the apples-and-oranges argument is dead.

Unfortunately, full meta-analysis is expensive in terms of needed data. For bare-bones meta-analysis, all you need is the study sample size and the study sample correlation. For full meta-analysis, you need measures of study quality on all relevant quality dimensions. At a minimum, you need the four basic measurement assessments: (*a*) the construct validity and reliability of the independent variable, and (*b*) the construct validity and reliability of the dependent variable. If studies differ in sampling bias, you need certain comparison standard deviations, and so on.

Unfortunately, current research practice is to pretend that your study is perfect. Researchers do not gather the needed data on study quality and hence do not report on study quality in their publications. Rather, the reviewer must search for the sporadic studies that do report study quality information. Under lenient assumptions, data on study quality can be sampled across the domain rather than required for all individual studies.

When quality data is only given sporadically, the estimates from meta-analysis are less accurate. To compensate for the lowered accuracy, you need more studies for a domain with imperfect studies than for a domain of perfect studies. That is, you need even more replications than you would need with statistical replications!

Summary on the Need for Replications

Scientific replications are more problematic than statistical replications. If studies are imperfect, then we need data measuring the extent of imperfection on the known dimensions where studies are imperfect. Thus we need more data for each study for imperfect studies than for perfect studies. At present, only a few researchers gather and report data on study quality; most researchers just blindly pretend that their study is perfect. Since the data on study quality is sporadic in most domains, the accuracy of meta-analysis on imperfect domains is greatly reduced. Two-digit accuracy for an imperfect domain will require even more replications than a domain of substantively perfect studies.

In conclusion, all current studies are imperfect. So if studies use different versions of the independent variable or use different versions of the dependent variable, then it is likely

that studies will differ in quality. Accurate determination of the size of the causal effect requires correcting the effects of such artifacts in order to eliminate spurious variation in observed effect sizes resulting from variation in study quality. The correction process not only requires extra information (such as the reliability and construct validity for each variable), but also larger sample size to reduce the additional sampling error introduced by error of measurement, biased sampling, and so on. So the earlier statements about sampling error made for pure statistical replications are underestimates of the number of replications needed. If the average sample size is as large as $N = 200$, you need more than 800 replications to achieve two-digit accuracy. The extent of increase depends on the average level of quality of the studies. For example, the multiplier for random error of measurement in the independent variable is the reciprocal of the reliability.

EXAMPLE:

Assume sample size needed for a perfect study is N_p.
Assume the study is perfect except for random error of measurement in the independent variable.
Reliability of the independent variable: r_{xx}.
Sample size needed for the imperfect study so that correction works as well as a perfect study: $N = (1/r_{xx})N_p$.

EXAMPLE:

$$N_p = 150{,}000 : r_{xx} = .80 : N = (1/.80)N_p$$

$$N = (1.25)(150{,}000) = 187{,}500.$$

We desperately need replication studies!

CONCEPTUAL REPLICATIONS

There is a kind of replication study that is easier to publish, though there is still resistance to publishing these studies too. The conceptual replication is a study that should be a replication but might not be. Laboratory researchers call the process of generating such studies the search for the missing control group. Any treatment, intervention, or manipulation is a bundle of administrative procedures, most of which are incidental to the active treatment ingredient.

When a researcher first introduces a treatment, there will be control groups to check the most obvious alternative hypotheses about administrative details that are thought to be irrelevant. The reader checks to see what hypotheses were not checked. The conceptual replication study then tests one such hypothesis. The experimental group is run with the incidental feature present, while the control group is run with the incidental feature absent. If the incidental feature is indeed incidental, then there will be no difference between the groups. If the original researcher was wrong, then the two groups will differ.

If the original researcher is right, then the new study will function as a replication of the original study. However, the one new element in the design may provide the needed creativity to get the study accepted.

There is little argument that conceptual replications are needed. However, a consideration of sampling error in such contexts shows that eventually a conceptual replication study will falsely find significance and will thus falsely assert that some incidental feature is actually critical. The only answer to that problem is to replicate the conceptual replication study. That replication of a replication study is likely to have trouble being published even though it is a perfectly well-done study. This is a bad practice. We desperately need replication studies. We need replication of main studies and replication of conceptual replication studies.

THE HETEROGENEOUS DOMAIN

If a domain is conceptually heterogeneous, then the true study population effect sizes differ across studies. This means the set of studies is not a set of perfect replications. There must be real differences between the studies. The reviewer who finds a heterogeneous domain can have one of three very different reactions to heterogeneity:

a) Panic: Eek! It's hopeless.

b) Measurement: How large is the departure from homogeneity?

c) Explanation: Why do results vary? Can the set of studies be split into subsets that are homogeneous?

The worst reaction to heterogeneity is panic. This is the position implicitly taken by the apples-and-oranges critics of meta-analysis. They say that since the studies are not perfect replications, the studies cannot be compared, and no meaningful analysis is possible. The data are then discarded!

Because of the way we use words, people are led to assume that there is an absolute distinction between homogeneity and heterogeneity. The distinction is actually fuzzy. I will present an example showing that the departure from homogeneity can be quantitatively trivial. This leads to the following question: Can we measure the extent of heterogeneity? As it happens, this is easy: we just look at the standard deviation of study population values. If that standard deviation is small, then the studies differ very little from one another. If that standard deviation is large, then the studies have substantial differences from one another. It is important to note that we can measure the extent of heterogeneity without understanding the causes of that heterogeneity.

Ultimately we will want to understand the heterogeneity. Why do study results differ from one another? Explanation of differences across studies requires substantive knowledge; it cannot be done by a simple statistical analysis. Statistics can help to test an explanation but cannot generate the explanation.

It should be noted that explanation of heterogeneity is greatly complicated by the fact that studies are imperfect for many reasons beyond just sampling error. Differences in study quality will produce differences in study outcome that have nothing to do with the issue being studied. That is, the true effect size could be uniform across studies, but the study outcomes could differ because of study imperfections.

Trivial Departures from Homogeneity

To the apples-and-oranges critics, any departure from perfect replication would result in aborting the review process. In particular, this would mean that any departure from homogeneity would mean that a domain should not be studied. The purpose of this example is to show that there can be a departure from homogeneity that is trivial in magnitude. In such a case, it would be disastrous to follow the advice of the apples-and-oranges critics.

Consider a set of 50 survey studies of the effect of stress on anxiety. Assume that for all studies, the population effect size for the ordinary correlation is $\rho = .30$. However, one study does not compute the ordinary correlation. In that study, the researchers artificially dichotomized the stress measure by splitting at the median. The population correlation for their study is thus reduced from $\rho = .30$ to $\rho = .24$. Since the results for this study differ from the other studies, the domain is not homogeneous but heterogeneous. The apples-and-oranges critics would reject any analysis of this domain of studies.

The fact is that the homogeneity significance test cannot detect this tiny departure from homogeneity. The test will register not significant and will thus falsely label the domain as homogeneous.

Would it be a major error to falsely conclude that the domain is homogeneous? Note first that there is no conceptual error in this conclusion. The deviant study outcome is not caused by a fact of nature but by the use of a deviant statistical analysis. If the deviant study had not dichotomized the stress measure, the study outcomes would have been perfectly homogeneous.

The point of this example is that even if a domain is not perfectly homogeneous, it is still possible for the variation to be so small as to be completely irrelevant. What we need is a measure of the size of the departure from homogeneity.

Measuring Heterogeneity

For a homogeneous domain, the population values are uniform across studies. For a heterogeneous domain, different studies have different population values. How do we measure the extent of departure from homogeneity?

The key to this is to think about the study outcomes as forming a distribution. That is, as we go across studies, we get various values for the population correlation. How would we describe that distribution? This is not a new question; we are just asking this question in a different context. Our usual answer is to describe the distribution by computing the mean as a measure of central tendency and by computing the standard deviation as a measure of spread from the mean.

Consider the mean and standard deviation in this context. We have a distribution of study population values ρ_i. We compute the mean correlation and the standard deviation of those correlations.

First, consider the homogeneous domain. If all population correlations are uniform in value, then there is a correlation ρ, and we have $\rho_i = \rho$ for all i. For this domain, the mean correlation will be the same as the uniform correlation. That is, the mean correlation will also be ρ in value. Since the values are uniform, the standard deviation will be zero.

Second, consider the heterogeneous domain where the values of the correlations ρ_i vary from study to study. If this distribution were approximately normal (not always a good approximation in this area), the mean correlation will be close to the middle of the set of study correlations. Since the correlations differ from each other, they will also differ from the mean correlation. The standard deviation will measure the extent to which study values differ from the mean study value.

Note that for a homogeneous domain, the standard deviation of population study values will be zero. For a heterogeneous domain, the standard deviation will be larger than zero. Furthermore, the larger the variation in study values, the larger the standard deviation. Thus the standard deviation of study outcomes is a very good measure of heterogeneity.

Consider the example of trivial heterogeneity above. Had the deviant study not dichotomized the stress measure, the mean correlation would have been .30 with a standard deviation of zero. With the deviant value, we have 49 studies with a correlation of .30 and one study with a correlation of .24. The mean correlation is thus only .2988 instead of .30. The standard deviation would be .0084 instead of zero. The description of effect sizes would then read as follows: for population effect sizes, Mean = .2988 and SD = .0084; for rounded to two digits, Mean = .30 and SD = .01.

For most scientific purposes, a standard deviation of .01 in relationship to a mean of .30 would have little theoretical or practical meaning. Nearly all inferences that simply assumed $\rho = .30$ would be entirely or substantially correct. Thus, the tiny standard deviation would show that the departure from homogeneity is trivial in magnitude.

Explaining Variation

Consider a set of studies that are believed to be scientific replications of each other. If a full meta-analysis finds the studies to be heterogeneous, then there is some error in the reviewer's beliefs. There is some aspect of the studies that matters when the reviewer would think that aspect to be irrelevant. However, if the standard deviation is very small, the error may be so small that it is hard to find the aspect.

If there is an existing theory that predicts differences in study results, then that theory should be tested. Indeed, a wise reviewer would know of this theory at the time of the review and would code the studies using that theory. The studies should be split using that code and compared to one another. If the mean correlation differs between the two groups of studies, then that not only shows that the domain is heterogeneous but also explains part of that variation in study outcome.

A study characteristic that causes differences in outcome is called a moderator variable. The size of the effect of a moderator shows up in the differences in effect size between

the different types of studies associated with different levels of the moderator variable. If all variation is to be explained by one moderator variable, then the size of the effect of the moderator variable is given by the standard deviation of population effect sizes in the meta-analysis.

In my experience, small moderator variables are usually found only if there is a preexisting theory that predicts variation in outcomes. If there is no such theory, it can be years before a moderator variable is identified.

If we have a large moderator variable, then the key question has to do with the number of studies in the domain. If the number of studies is large, we get a good test of any alleged moderator variable. If the number of studies is small, then it is hard to test a potential moderator variable. For example, the accuracy of a significance test depends critically on the number of studies. The fewer the studies, the higher the error rate for the significance test.

If the number of studies in the domain is small, then there is a major problem in blindly looking for a moderator variable: the problem of capitalization on chance. If we define a large number of potential moderator variables by just checking any study characteristic that can be coded, then we provide a large number of opportunities for the significance test to fail and generate a false claim for a moderator variable.

When a significance test fails and falsely suggests that some study characteristic is a moderator variable, there will be a difference between the mean effect sizes for the subsets defined by that characteristic. That difference is actually because of sampling error. The size of that difference will depend on the number of studies in each subset. The smallest subset is likely to be the one with the large sampling error that caused the significance test to fail. The size of that error depends on the total sample size in that smallest subset, and that depends on the number of studies in the smallest subset.

If we need a moderator analysis to understand a heterogeneous domain, then the consideration of sampling error shifts from the total domain to the subsets defined by the moderator variable. If we need 800 studies to get two-digit accuracy in a homogeneous domain, then we need 800 studies in the smallest subset to get accuracy within each study type.

Consider the best-case scenario. The moderator variable is a binary variable, and so we only have two subsets. The best case for the smallest subset comes about if the frequency for the two types of studies is the same. That is, if the domain has a 50-50 split between study types, then the subsets each have half the studies, and the number of studies within types is exactly half the total number of studies. If we need 800 replications for two-digit accuracy in estimating the effect size, then we need 800 replications of each type for a total number of 1,600 replications.

If we have a binary moderator variable with an uneven split, the situation can be much worse. Suppose we have an 80-20 split between types. If we need 800 replications for the small subset, then we will have 3,200 replications for the large subset and thus 4,000 replications altogether. If the moderator variable takes on more than two values, then the situation can be very bad because the smallest subset may be quite smaller than the other subsets, and there may be considerable sampling error in estimating the average effect size for that subset.

GENERAL CONCLUSION

Many journals base publication decisions almost entirely on the creativity of the article. Replication studies of any kind are then regarded as useless and are thus hard to publish. I argue in this essay that such a position is disastrous science. In science, facts are at least as important as ideas. Replication studies are desperately needed in order to determine facts.

Consider a domain of substantively perfect studies where sampling error is the only problem. If the effect size is measured by a correlation, the sample size needed for two-digit accuracy is $N = 150{,}000$. Even in a domain with relatively large samples such as consumer research, this is a massive sample size for single studies. The average sample size is about $N = 200$, and it would take 750 studies to have the needed total sample size. That is, in a domain of perfect studies with relatively large sample sizes, we need a minimum of nearly 800 replication studies. These studies would be pure replications—no creativity at all.

In other areas of research, the situation is even worse. The average sample size in overall psychology is about $N = 100$. For intense individual treatments (such as therapy or neural recording) in psychology and medicine, the average sample size is often $N = 20$ or less. For substantively perfect studies with an average sample size of $N = 20$, we need 7,500 replication studies to get accuracy to two digits. That is, there are many important areas in social and biological science where we need thousands of replication studies in order to get accurate statements of effect size.

Since there are no perfect studies in real science, the figures above present a rosy scenario view of our database. If studies are imperfect, we need information measuring the size and extent of these imperfections. Hunter and Schmidt (1990) have reviewed the facts about 13 known dimensions along which studies can be imperfect (i.e., our legacy from psychometric theory). More dimensions of imperfection will be identified in the future. Given the right information, the distortions in the data produced by imperfections can be corrected. But correction comes at a price. To correct an imperfection, we need measurement of the size of that imperfection (e.g., random error is measured by the reliability coefficient), which is new and often unreported information. Furthermore, when we correct an imperfection, the standard error of the corrected correlation will be considerably larger than the standard error for a perfect study. So imperfect studies need much larger sample sizes than perfect studies because the correction process is required. Thus for imperfect studies, we need even more replications than the large number required to reduce sampling error in perfect studies. For example, in a domain with a relatively high reliability of .80 for both the independent and dependent variable, the

sample size multiplier for random error of measurement would be (1/.80)(1/.80) = (1/.64) = 1.56. That is, even in studies with very good measurement by current standards, the sample size needed for correction would be 56 percent larger than the sample size for a study with perfect measurement.

We desperately need replication studies, whether creative or not. Furthermore, we do not just need one replication study, we need many replications. For a rough estimation, we need 10 replication studies. In the long run, we will need hundreds of replications in large sample domains and thousands of replication studies in small sample domains.

[*David Glen Mick served as editor for this article*]

REFERENCES

Hunter, John E. and Frank L. Schmidt (1990), *Methods of Meta-analysis*, Thousand Oaks, CA: Sage.

Sedlmeier, Peter and Gerd Gigerenzer (1989), "Do Studies of Statistical Power Have an Effect on the Power of Studies?" *Psychological Bulletin*, 105 (March), 309–316.