Acad. Quest.
DOI 10.1007/s12129-019-09778-5

ARTICLES

Object of Inquiry: Psychology's Other (Non-replication) Problem

John Staddon

Ever since the 2005 publication of a landmark paper "Why Most Published Research Findings Are False" by medical statistician John Ioannidis,[1] social and biomedical science has been stumbling through what is now termed the "Irreproducibility Crisis."[2] The "crisis" refers to the failure to get the same results when supposedly reliable, "statistically significant" studies are repeated.

The crisis is a legacy of what has become the dominant experimental method in social and biomedical science, establishing a true effect by comparing two or more *groups* of subjects each exposed to a different experimental manipulation. In this article I argue that remedies proposed for this problem miss an important point. The problem is not just the statistical reliability—repeatability—of group results, which is already being improved. An equally important problem is drawing conclusions about the psychology or physiology of individuals from group averages.

The aim of basic research in biomedicine and social science is to understand the psychology or physiology of individual human beings. But more applied areas of research, such as polling, education, or agriculture, are interested in the effects of a treatment on groups—populations. They are concerned with

[1]John P. A. Ioannidis, "Why Most Published Research Findings Are False," *PLoS Med* 2, no. 8 (2005), https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1182327/; *Reproducibility and Replicability in Science*, National Academies of Science, Washington D. C, https://www8.nationalacademies.org/pa/projectview.aspx?key=49906;
G. Gigerenzer, (Statistical rituals: The replication delusion and how we got there," *Psychological Science* 1, no. 2018: 198-218; *The Irreproducibility Crisis of Modern Science,* National Association of Scholars, New York, https://www.nas.org/images/documents/NAS_irreproducibilityReport.pdf

[2]M. Baker, "1,500 scientists lift the lid on reproducibility: Survey sheds light on the 'crisis' rocking research," *Nature*, https://www.nature.com/news/1-500-scientists-lift-the-lid-on-reproducibility-1.19970.

John Staddon is James B. Duke Professor Emeritus, Department of Psychology and Neuroscience, Duke University.

Published online: 05 April 2019

individual differences only to the extent that they affect group differences. These different objects of study have strong implications for the appropriate research methods. I begin with examples of both types of research.

Research on Individuals: The Single-Subject Method

Experimental psychologists are interested in topics ranging from vision and perception to learning and choice behavior. Some phenomena can be studied one subject at a time. For example, a test for color perception, such as the famous Ishihara test, can be given repeatedly to the same person to see if the diagnosis is reliable. The result of the test is essentially independent of intervening experience. The overhead light can go off and on, or change color in between tests, and providing the person is not dazzled, the test will always give the same result. Many problems in perception, and even adaptation to reward schedules, are like that. The same stimulus or schedule will give the same response. Only a handful of subjects, studied one at a time but tested repeatedly, are needed, just to ensure that the demonstrated effect is not an oddity. Statistics are unnecessary.

But many phenomena cannot be repeated with any assurance that the same treatment will give the same result. By its very nature, data that involves *learning* does not bear repeating. A classic learning problem is: which is better, massed or spaced learning? Should to-be-learned items (nonsense syllables, say) be spaced close together in time or far apart? Which yields the faster learning? To answer this problem with a single subject is obviously tricky because whichever spacing is learned first will have an effect on whichever is learned second. There is no "reset" button that can restore the subject to a neutral state in between procedures. Nevertheless, Hermann Ebbinghaus (1850-1909), a pioneer who managed to discover many of the basic phenomena of verbal memory, used only himself as a subject. Ebbinghaus's success was not sufficient to prevent abandonment of the single-subject method by the field that he founded. Indeed, someone who tried to publish today a memory study using Ebbinghaus's strategy would see his submission rejected out of hand.

The contradiction between Ebbinghaus's success with a single subject and the rejection of his method by modern memory researchers is a paradox that has attracted little attention.[3] The reason for this lack of interest may be that Ebbinghaus's solution to the problems posed by the irreversibility of learning involved considerable ingenuity. But the alternative method, which I discuss next,

[3]Despite a lengthy treatment by social psychologist and historian Kurt Danziger, *Constructing the Subject* (Cambridge, MA: Cambridge University Press, 1990).

is much simpler: just follow a well-defined set of "gold standard" rules. This apparent simplicity has made misuse of the method all too easy, as we will see.

Research on Groups: The Between-Subjects Method

The usual solution to the problem of studying irreversible or only partially reversible phenomena, has been the *between-subjects* experimental method, which is now almost universal in biomedicine and social science. For example, suppose an educational psychologist wishes to compare two ways of teaching children to read. Clearly, he must choose one or the other method to present to a given child. Absent a "reset" button, he can't train a child one way and then the other and compare the results. An intermediate example is human choice behavior, which I will say more about in a moment.

Ever since the groundbreaking work of English statistician and biologist R. A. Fisher (1890-1962) the usual way to deal with problems like this has been to compare the two treatments not in one individual but between two groups.[4] The group method has become standard. In the education example, children are chosen from a larger group and assigned randomly to one of the two treatment groups. Randomness is important, to ensure that the two groups do not differ for some accidental reason—e.g., all the smart kids are in one group, or one group has more dyslexic kids than the other. And the groups must be relatively large, for the same reason: to ensure that the unavoidable individual differences among the children will cancel out in the average.

After the treatment, the results—the average performance of each group and the variability of performance (variance) within each group—must be compared. Sometimes this is easy. If the two groups are large and the data don't overlap at all, no further analysis is necessary. If no child in Group A does better than the worst child in Group B, clearly treatment B is better. But if the data overlap, some other way must be found to evaluate the two treatments. To do this in a valid way requires (and this is often forgotten) a *model*, a hypothetical process that explains the variability in the data. Studies that use standard statistics to analyze their data rarely explain or even justify the underlying statistical model. Possible models and their properties are topics of a large and highly technical statistical literature that I have discussed elsewhere.[5] For the moment, just note that the big unknown is the source(s) of variability.

[4]E. B. Ford, R. A. Fisher, "An appreciation," *Genetics*, 171(2005), 415-41; J. F. Box *R. A. Fisher: The life of a scientist* (Wiley, 1978).

[5]John Staddon, *Scientific Method: How science works, fails to work or pretends to work.* (Routledge: Taylor and Francis, 1978).

An early application of statistics treated errors in astronomical measurements, by Belgian astronomer and polymath Adolph Quetelet (1796-1874).[6] Physical vibration and changes in atmospheric refraction cause small differences in repeated measurements of the same astronomical quantity—the exact location of a star, for example. It was perfectly reasonable to divide the measured value, V_M, say, into two parts: the (hypothetical) real value, V_R which is assumed to be fixed and an added "error" term, e: $V_M = V_R + e$; e is usually assumed to be drawn from a random "normal" distribution with mean zero.

The assumption of additive, "normal" variability in this situation is both plausible and mathematically tractable. It is now applied almost universally, even to situations much more complex than physical measurement—situations where the target variable may not in fact be constant, and where there may be multiple sources of variability, many of them non-additive and non-random.

As we will see, Quetelet's aim differed from Fisher's. Fisher, whose seminal work was done at the Rothamsted Agricultural Station in England, was concerned with the factors that affect the fertility of a field: rainfall, temperature, fertilizer etc. The between-group method in his words was "the study of **populations**":

> From a limited experience, for example, of individuals of a species, or of the weather of a locality, we may obtain some idea of the *infinite hypothetical population* from which our sample is drawn and so of the probable nature of future samples to which our conclusions are to be applied . . .
>
> The salutary habit of repeating important experiments, or of carrying out original observations in replicate, shows a tacit appreciation of the fact that *the object of our study is not the individual result,* but the *population of possibilities* of which we do our best to make our experiments representative. The calculation of means and probable errors shows a deliberate attempt to learn something about that population. [emphasis added]

And finally:

> [I]n most cases only an infinite population can exhibit accurately, and in their true proportion, the whole of the possibilities arising from the causes actually at work, and which we wish to study.[7]

[6] G. Jahoda, "Quetelet and the emergence of the behavioral sciences." Published online, September 2015: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4559562/

[7] R. A. Fisher, *Statistical Methods for Research Workers.* (Edinburgh, Oliver & Boyd: 1925), http://www.haghish.com/resources/materials/Statistical_Methods_for_Research_Workers.pdf

Fisher's concern was very different from Quetelet's. Quetelet was not interested in an infinite population. He wanted to find a single true value for a measurement that was blurred by what we would now call "noise." That accurate number could then be compared with a theoretical prediction or form part of a database that might lead to new theory.

Although Fisher did not say so explicitly, his agricultural work was not about basic science, not about understanding how some treatment, fertilizer, or soil type, say, affects plant physiology. It was applied science, about making practical decisions. His interest was not in a particular, randomly assigned plot, but on the overall effect of a treatment on a sample of such plots. Which treatment is better? These two concerns, Quetelet's and Fisher's, are very different: the one aims at certainty, a single true value for a measurement; the other accepts variability as inevitable and deals with it as best it can in the service of practical action. In much of social science and biomedicine, these two very different concerns are now confused.

Choice Behavior

Fisher's concern with populations and the methods he introduced to deal with them are in fact ill-suited to basic psychological science. A famous example from the study of human choice behavior shows what I mean. People will usually respond immediately to hypothetical choice questions, such as "Do you prefer $2,000 with a probability of 60 percent return or $1,000 guaranteed?"—which allows experimenters who use this technique to test hypotheses very quickly. This strategy was adopted by Daniel Kahneman and the late Amos Tversky in a groundbreaking—and Nobel-prize-winning—series of papers beginning in the 1970s. Their work provided an important impetus for what has become known as *behavioral economics*.[8]

Kahneman and Tversky posed questions to themselves and checked their own answers later with groups of human subjects. Their results are statistically significant, and many have in fact been replicated. They were able to show that people answered their questions in ways that violated the standard version of utility theory, the usual outcome-based explanation in economics.

Swiss mathematician Daniel Bernoulli, discoverer of (among many other things) the principle that allows airplanes to fly, pointed out in the eighteenth century that the value (utility) of a given increment in wealth (goods or cash) is

[8]"Behavioral Economics," Wikipedia.com, https://en.wikipedia.org/wiki/Behavioral_economics; J.E.R. Staddon, *Limits to action: The allocation of individual behavior* (New York: Academic Press, 1980) for similar approaches to learning in animals.

inversely related to the amount of wealth you already have. A hundred dollars is worth much more to a pauper (net wealth: $5) than to a millionaire. Bernoulli proposed that the relationship between wealth and utility is logarithmic, so that equal ratios correspond to equal value. A pay raise of $1,000 is worth as much to an employee whose base pay is $10,000 as a raise of $10,000 to one who makes $100,000. Bernoulli's hypothesis means that the utility function for any good is curved—negatively accelerated, or *concave*, in the jargon. In other words, Bernoulli's utility function shows that old favorite: diminishing marginal utility. It is the standard, not just for money, but for almost all goods.

The concave utility curve predicts that (rational!) people will tend to be risk averse with respect to gains. The reason? Each increment of a good adds a smaller and smaller increment of utility; hence, doubling the amount of a good less than doubles its utility. Which means that a person might well prefer 100 percent chance of X to a 60 percent chance of 2X: $100 with probability 1.0 over $200 with probability 0.6.

And in Kahneman and Tversky's experiments most people did. In one experiment, 95 people were asked to choose between two outcomes: $4,000 with probability 0.8, vs. $3,000 guaranteed. Eighty percent (note: *not* 100 percent) chose the sure thing, even though it is less than the expected value of the gamble: $3,000 < $3,200 . Apparently, for most people 0.8 times the utility of $4,000 is less than the utility of $3,000, 0.8xU(4000) < U(3000)—because the utility function is concave. an 80 percent chance of $4,000 vs. $3,000 guaranteed.

Other experiments showed that things are not so simple; and Bernoulli's account is what is called *teleological*, i.e., based on a goal or outcome, not causal, as are most explanations in the "hard sciences." But the point I want to make here is that in *none* of these experiments was preference unanimous. Sometimes as few as 65 percent of subjects chose the majority option. It is emphatically not the case that every human being in every circumstance shows risk sensitivity or (another term from this line of work) confirmation bias. The group data were replicable; individual-subject data were not. It is this kind of uncertainty that seemed to make statistical proof essential.

The problem—a problem that has been largely ignored in the debate about irreproducibility—is that the object of Kahneman and Tversky's inquiry is not and cannot be the individual decision maker. The object of their inquiry is the population of individuals from which their rather arbitrary sample of subjects was drawn. This is a limitation of these studies as contributions to psychology—but not to economics. It is no wonder, therefore, that they led to further developments in economics rather than to advances in our understanding of the mechanisms of individual choice.

When individual data show exceptions, what is the alternative to statistical analysis of groups?

Back to the Single Subject

The alternative is to abandon statistics and averages as the endpoint of any inquiry whose aim is the understanding of individual organisms. The single subject method, whether the subject is a chemical, an electromagnetic circuit, or even a human being, long preceded R. A. Fisher.

The standard for biological and human research was set by the father of experimental medicine, Claude Bernard (1812-1878) in his seminal work *An Introduction to the Study of Experimental Medicine* (1865).[9] Fisher came after Bernard, and his statistics are more sophisticated than those with which Bernard was familiar. But sophisticated or not, all statistics rest on the same foundation: treating the data as samples of a population rather than properties of an individual. Bernard was concerned with individuals and made a strong claim:

> [T]hat the word *exception* is unscientific; and as soon as laws are known, no exception indeed can exist, and this expression, like so many others, merely enables us to speak of things whose causation we do not know.

Bernard was the discover of the constancy of the *milieu intérieur* (biological homeostasis), the idea that a multitude of physiological variables are regulated and that this regulation is essential to life. He hated averages and thought them useless as a guide to the functioning of the body:

> In every science, we must recognize two classes of phenomena, first, those whose cause is already defined; next, those whose cause is still undefined. With phenomena whose cause is defined, statistics have nothing to do; they would even be absurd. As soon as the circumstances of an experiment are well known, we stop gathering statistics: we should not gather cases to learn how often water is made of oxygen and hydrogen; or when cutting the sciatic nerve, to learn how often the muscles to which it leads will be paralyzed. The effect will occur always without exception, because the cause of the phenomena is accurately defined. Only when a phenomenon includes conditions *as yet undefined*, can we compile statistics; we must learn, therefore, that we compile statistics only when we cannot possibly

[9] John Staddon, "Claude Bernard's Introduction to the Study of Experimental Medicine," *The New Behaviorism*, October 28, 2018, https://sites.duke.edu/behavior/2018/10/28/claude-bernards/

> help it; for in my opinion *statistics can never yield scientific truth*, *and therefore cannot establish any final scientific method*.

He goes on to give an example:

> Certain experimenters, as we shall later see, published experiments by which they found that the anterior spinal roots are insensitive; other experimenters published experiments by which they found that the same roots were sensitive. These cases seemed as comparable as possible; here was the same operation done by the same method on the same spinal roots. Should we therefore have counted the positive and negative cases and said: the law is that anterior roots are sensitive, for instance, 25 times out of a 100? Or should we have admitted, according to the theory called the law of large numbers, that in an immense number of experiments we should find the roots equally often sensitive and insensitive? Such statistics would be ridiculous, for there is a reason for the roots being insensitive and another reason for their being sensitive; *this reason had to be defined*; I looked for it, and I found it; so that we can now say: the spinal roots are always sensitive in given conditions, and always insensitive in other equally definite conditions. [emphasis added]

Bernard has a simple message for a scientific community grappling with a replication crisis. The problem is not that a particular group result cannot be reliably repeated. That can be solved by pre-registering hypotheses, and using larger groups and more conservative tests of statistical significance. (All these changes have occurred in the aftermath of Ioannidis's revelations.) The group is the object of inquiry, and a suitably designed experiment will be 100 percent replicable—at the group level. The problem is that even if a result, like Kahneman and Tversky's group choice data, can be reliably repeated, it still tells us nothing about the individual. To say that a given set of subjects, drawn from a population collected for convenience, makes the risk-averse choice 80 percent of the time tells us as much about the choice process in individual human beings as a statistic about birth rates tells us about why a given couple has three rather than one or no children. To say that human beings are risk averse, always and everywhere, based on a finding from a fraction of people drawn non-randomly from the population at large is, Claude Bernard would surely say, *absurde*.

So, what is the solution? Why do people differ in their answers to these choice questions, albeit in a biased way (80 percent choosing one way, only 20 percent the other, etc.)?[10] The answer is the same as Bernard's question about differences in spinal root sensitivity: what is different about a person who is risk averse in this situation as opposed to one who is not? Is it their constitution, their genetic makeup? If we ask the same question repeatedly, do they always give the same answer, no matter what? Do they answer similar but differently phrased questions differently? Not "Which would you choose?" but "How would an economist/statistician/child choose?" Or "Which one pays off the most?" Does the same subject answer the question differently under different circumstances or are individual differences stable? Or do individual differences reflect the different personal histories of the subjects? Is Jeb late paying his bills? Has Alice just won a lottery? Is Joan a Democrat or Republican? How would Kahneman and Tversky have answered their own questions?

Explorations like these, following the tortuous path necessary to understand the process that underlies choice behavior, are much more difficult than simply asking the same question of 95 people. The search for invariance requires great ingenuity; no algorithmic "Gold Standard," is available. It is real science. But it is essential to understanding the process, the invariant property or properties of human beings that causes them to respond in certain ways in certain situations. Understanding these processes is the proper aim of both psychological and biomedical science.

Statistical results from group experiments can provide insights into human behavior *no higher than the level of public opinion polling*. Indeed, experiments like Kahneman and Tversky's, involving no manipulation more drastic than simply asking a question or two of a haphazardly chosen subject population, are conceptually identical to opinion polling. The only differences are the number of questions, the rule by which the subject population is selected, and the time chosen for the exercise.

The unavoidable conclusion is that the brilliant statistical methods of R. A. Fisher and his collaborators and successors are simply inappropriate for questions of individual psychology. If understanding the individual organism is the aim, the solution to the replication crisis is not to accommodate Fisher's statistics but to abandon them. To understand how the mind/brain works, we need to study the psychology and neurophysiology of individuals with the same ingenuity and tenacity that Bernard applied to their physiology. Only when psychology can point to individual certainties will it have attained the status of a science.

[10]Gerd Gigerenzer raised questions like this: "How to Make Cognitive Illusions Disappear: Beyond 'Heuristics and Biases,' in W. Stroebe, M. Hewstone *European Review of Social Psychology* 2 (1991): 83–115.

What is to be done?

My aim is not to discredit Fisherian methods. They work and will remain useful if two things are true: the object of study is a population and not individuals, and if the aim of the study is to make a practical decision. In Fisher's case the decision involved treatments like fertilizer applied to fields. The aim of the experiment was to decide which fertilizer, A or B, was better. The agriculturalists were not concerned with why it was better. Nor were they concerned with cost. A full analysis should conclude not just that A is a better fertilizer, but that it was better by enough of a margin to compensate for its added cost. But all in all, Fisher's approach is appropriate for the subjects to which he applied it. It offers a way to decide between different public health policies, for example. Is vaccinating better than not vaccinating? Are the risks of vaccination outweighed by its benefits?[11] Is vaccine mix A better than mix B? Issues like this can be settled by comparing suitably chosen control and experimental groups.

A scan of scientific journals reveals many similar examples: "University-affiliated alcohol marketing enhances the incentive salience of alcohol cues" deals with the effect of signage on underage alcohol consumption.[12] The target is an undergraduate population, not the psychological processes of individuals. The paper informs policy for university administrations. The Fisherian apparatus of group comparison and significance tests is an appropriate endpoint.

On the other hand, another article in the same journal, using the same methods, comes to conclusions about individual psychology.[13] The researchers gave both members of a couple a computerized test and then looked to see how they interacted. They found that the test results predicted aspects of their nonverbal interaction:

> The . . . research [showed] . . . that participants with more positive implicit partner evaluations exhibited more constructive nonverbal (but not verbal) behavior toward their partner in a videotaped dyadic interaction . . . These findings represent a significant step forward in understanding the crucial role of automatic processes in romantic relationships.

[11] A. G. Walton, "A cost-benefit analysis of vaccines," *The Atlantic*, January 23, 2012, ttps://www.theatlantic.com/health/archive/2012/01/a-cost-benefit-analysis-of-vaccines/251565/

[12] Barthelow et al. "University-Affiliated Alcohol Marketing Enhances the Incentive Salience of Alcohol Cues," *Psychological Science* 29, no 1 (2017): 83-94.

[13] Faure et al., "Speech is silver, nonverbal behavior is gold: how implicit partner evaluations affect dyadic interactions in close relationships," *Psychological Science* 29, no. 11 (2018): 1731-1741.

These findings were established via the usual apparatus of significance tests. The numbers were large (129 subject-pairs) and significance levels were also high ($p < .001$, usually). The study has not been replicated, but the large number of subjects and high significance levels suggest that the group results are probably reliable. The study meets the objections that have been quite properly raised by critics of irreproducibility.

But of course, many individual pairs did not show the reported effect. There were exceptions to the reported conclusion. Why? What was different about the couples who failed to show a correlation between "implicit evaluations" (measured by an association test) and "more constructive nonverbal behavior" towards their partner in a later test session.[14] Were they less "happy"? Did they have reliably different histories? Were they different personality types? Would individual couples respond in the same way if the experiment were repeated?

The current concern about the replicability of group comparison science is legitimate. But it is only half the problem. The other half is the plethora of replicable group studies that make claims about the psychology, the mind/brain, of individual subjects. Editors should give considerably greater care to the object of inquiry in the study under their review. Does the study make claims about a group, like the alcohol marketing study? Is it ambiguous, both about group behavior (economics) and individual psychology, like the Kahneman and Tversky experiments? Or is it about individual subjects (subject pairs), as in the implicit evaluation study? In the last case, probably the proper course is to ask for more research: few, if any, scientific papers that draw conclusions about individual psychology based solely on group data should ever reach publication. Only when there are no exceptions, and significance tests are no longer required, can we be satisfied that there has been real advance in understanding the psychology of individual human beings.

[14] A. Karpinski, R. B. Steinman, R. B., "The Single Category Implicit Association Test as a measure of implicit social cognition," *Journal of Personality and Social Psychology* 91 (2016): 16–32.