What Type of Type I Error? Contrasting the Neyman-Pearson and Fisherian Approaches in the Context of Exact and Direct Replications

Mark Rubin
The University of Newcastle, Australia

Citation: Rubin, M. (2019). What type of Type I error? Contrasting the Neyman-Pearson and Fisherian approaches in the context of exact and direct replications. *Synthese*. https://doi.org/10.1007/s11229-019-02433-0

Abstract

The replication crisis has caused researchers to distinguish between *exact replications*, which duplicate all aspects of a study that could potentially affect the results, and *direct replications*, which duplicate only those aspects of the study that are thought to be theoretically essential to reproduce the original effect. The replication crisis has also prompted researchers to think more carefully about the possibility of making Type I errors when rejecting null hypotheses. In this context, the present article considers the utility of two types of Type I error probability: the Neyman-Pearson long run Type I error rate and the Fisherian sample-specific Type I error probability. It is argued that the Neyman-Pearson Type I error rate is inapplicable in social science because it refers to a long run of exact replications, and social science deals with irreversible units (people, social groups, and social systems) that make exact replications impossible. Instead, the Fisherian sample-specific Type I error probability is recommended as a more meaningful way to conceptualize false positive results in social science because it can be applied to each sample-specific decision about rejecting the same substantive null hypothesis in a series of direct replications. It is concluded that the replication crisis may be partly (not wholly) due to researchers' unrealistic expectations about replicability based on their consideration of the Neyman-Pearson Type I error rate across a long run of exact replications.

Keywords: direct replication; exact replication; Fisher; Neyman-Pearson; Type I error; replication crisis

 Correspondence concerning this article should be addressed to Mark Rubin at the School of Psychology, Behavioural Sciences Building, The University of Newcastle, Callaghan, NSW 2308, Australia. Tel: +61 (0)2 4921 6706. E-mail: Mark.Rubin@newcastle.edu.au Web: http://bit.ly/QgpV4O

In 2015, the Open Science Collaboration attempted to replicate the findings of 100 psychology studies. They found that only "39% of effects were subjectively rated to have replicated the original result" (Open Science Collaboration, 2015, p. 943). More recently, a group of researchers attempted to replicate the findings of 21 social science studies published in *Nature* and *Science* between 2010 and 2015 (Camerer et al., 2018). Using relatively large sample sizes, this project found that 62% of studies replicated the original results. Finally, another international multi-lab replication attempt of 28 psychology effects found that 54% replicated (Klein et al., 2018). These replication rates are generally considered to be unsatisfactory, and they have contributed to a *replication crisis* in the social sciences and beyond (for a review, see Shrout & Rodgers, 2018).

There have been two main responses to the replication crisis. The first response has been to focus on the statistical practice of hypothesis testing, with a particular emphasis on *p* values and Type I error probabilities (e.g., Benjamin et al., 2018; Lakens et al., 2018). This approach aims to help researchers to better distinguish between potentially replicable effects and nonreplicable noise.

The second response has been to encourage more replication attempts in order to better understand which effects are actually replicable and which are not (e.g., LeBel, Berger, Campbell, & Loving, 2017; Zwaan, Etz, Lucas, & Donnellan, 2018). This response involves making original research studies more reproducible by making research materials openly available to other researchers and by facilitating the publication of replication attempts (Nosek, Spies, & Motyl, 2012).

In the present article, I provide an integrative discussion of these two responses to the replication crisis by considering two different types of replication in relation to two different types of Type I error probability. Specifically, I consider the distinction between *exact* and *direct* replications and highlight the point that exact replications are impossible in social science, whereas direct replications are possible and essential. I then consider the Neyman-Pearson long run Type I error rate and argue that it is only meaningful in situations in which a long run of exact replications are possible. Given that exact replications are impossible in social science, I argue that the Neyman-Pearson Type I error rate is not meaningful in social science. Instead, the Fisherian Type I error probability is more appropriate because it can be implemented in situations in which exact replications are impossible. Hence, I argue that the Fisherian Type I error probability is more applicable in social science than the Neyman-Pearson Type I error rate because it does not rely on the concept of a long run of exact replications.

I should note that some commentators argue that researchers should abandon significance testing and, with it, the concept of Type I errors (e.g., Amrhein, Greenland, & McShane, 2019; Wasserstein, Schirm, & Lazar, 2019). In the present paper, I assume that readers are interested in undertaking significance testing, and I address the second-order question of which significance testing approach is more appropriate in the context of the replication crisis: the Neyman-Pearson approach or the Fisherian approach.

I should also note that I do not focus on the Bayesian approach to hypothesis testing in this article because many articles have already compared significance testing with Bayesian hypothesis testing (e.g., Berger, 2003; Berk, Western, & Weiss, 1995; Fisher, 1959, p. 17, 20-23; Wagenmakers, 2007; Wagenmakers et al., 2018; Wagenmakers & Gronau, 2018). Critically, these articles tend to compare the Bayesian approach with the Neyman-Pearson approach, rather than the Fisherian approach.[1] In the current article, I limit my considerations to significance testing, and I compare the Neyman-Pearson and Fisherian approaches in the context of exact and direct

replication studies. To my knowledge, no previous articles have addressed this particular issue, and yet it is one that has important implications for scientists who use significance testing. In particular, it affects the meaning and interpretation of "statistically significant results" as well as expectations regarding the replication of these results. I begin with a comparison of exact and direct replications.

Different Types of Replication

Exact Replications are not Possible in Social Science

Exact replications require the duplication of *all* of the aspects of an original study that could potentially affect the results of that study. These aspects include the sampling procedure, sample size, testing conditions, stimuli, measures, data coding and aggregation method, and analyses (e.g., Lindsay & Ehrenberg, 1993; Schmidt, 2009; Shrout & Rodgers, 2018). In social science, each of these methodological aspects is likely to vary from one study to the next. For example, consider a study that investigates the hypothesis that men have higher self-esteem than women. If an initial study samples participants from the U.S.A., and a replication attempt samples participants from France, then the sampling procedure has changed. If the initial study measures self-esteem after measuring the perceived societal status of men and women, but the replication attempt does not, then the testing conditions have changed. If the initial study uses Rosenberg's (1965) Self-Esteem Scale, and the replication attempt uses a French translation of this scale, then the measurement approach has changed. Finally, if the initial study controls for age in its analysis, but the replication attempt does not, then the analytical approach has changed.

Even if all methodological aspects of a study are kept exactly identical from one sample to the next, participants' culturally-based interpretation of the method is likely to undergo systematic changes across time and location (Billig, 2018; Cesario, 2014; Iso-Ahola, 2017; Schmidt, 2009; Schwarz & Strack, 2014; Stroebe & Strack, 2014; Zwaan et al., 2018). For example, participants' interpretation of the items in Rosenberg's (1965) Self-Esteem Scale is likely to be different depending on whether they were born in 1970 or 2020 and depending on whether they grew up in the U.S.A. or China. Consequently, as Serlin (1987) explained, "there is no psychological basis for expecting conclusions to hold for a population that differs in any respect from the sampled one, including the population into which the sampled population evolves an hour after sampling" (p. 366; see also Earp & Trafimow, 2015, p. 3; Schmidt, 2009, p. 92). The Greek philosopher Heraclitus put it this way: "No man ever steps in the same river twice, for it's not the same river and he's not the same man." Of course, the river and man in question retain many stable features from one moment to the next. Nonetheless, they will also undergo changes that distinguish them from their past versions. In this sense, rivers and people are what Schmidt (2009, p. 92) called *irreversible units* in that they are complex time-sensitive systems that accumulate history. The scientific investigation of these irreversible units cannot proceed on the assumption that exact replications are possible. Social scientists need to take into account the fact that people are time- and context-sensitive units of analysis that have the potential to interpret identical situations in multiple different ways (e.g., Ferguson, Carter, & Hassin, 2014), and they need to interpret their research results as being the product of an "interaction between general processes and the social context in which they operate" (Tajfel, 1981, p. 21; see also Billig, 2018). Consistent with this context-based interpretation, a reanalysis of the Open Science Collaboration's (2015) psychology replication attempts found that effects that were more likely to be contextually sensitive (i.e., more likely to vary in time, culture, or location) were less likely to be replicated (Van Bavel, Mende-Siedlecki, Brady, & Reinero, 2016; cf. Klein et al., 2018; Stanley, Carter, & Doucouliagos, 2018).

Researchers can eliminate the influence of methodological and cultural variations by randomly dividing their sample into two or more subsamples and then attempting to replicate their results from one subsample to the next. However, even this *holdout subsample* approach does not produce exact replications because, unlike the parent sample, the subsamples are not randomly sampled from the population. They are only randomly sampled from the parent sample. Consequently, although subsamples allow exact replications to be conducted with respect to the parent sample, they do not allow exact replications to be conducted with respect to new samples that are randomly drawn from the population (Krause, 2019, Footnote 1).

Finally, even if exact replications were possible in social science, they would not be meaningful in some cases because many social effects have sociocultural causes that vary over time and location, causing corresponding changes in their size and even their existence (Billig, 2018; De Boeck & Jeon, 2018; Ferguson et al., 2014, p. 301; Iso-Ahola, 2017; Strack, 2017). To illustrate, Zuckerman, Li, and Hall (2016) conducted a meta-analysis of the effect in which men report higher self-esteem than women. They found that this "gender difference emerged after the 1970s, increased until 1995, and declined afterwards" (Zuckerman et al., 2016, p. 34). They proposed a historical model as a potential explanation of these changes. Assuming that this historical model is correct, failure to demonstrate a gender difference in self-esteem in an exact replication in 2020 would not be inconsistent with the demonstration of this gender difference in 1995 because the size of the true effect is assumed to have declined.

Taken together, the above issues have led several researchers to conclude that exact replications are not possible or useful in social science (Anderson et al., 2016; Berk & Freedman, 2003; Berk et al., 1995; Brunswik, 1955; Camilli, 1990, p. 137; Cumming, 2008; De Ruiter, 2018; Earp & Trafimow, 2015; Hampel, 2003, p. 3; Hansen, 2011; Iso-Ahola, 2017; Lindsay & Ehrenberg, 1993; Macdonald, 1997, p. 337; Nosek & Errington, 2017, 2019; Nosek et al., 2012; Rubin, 2017a; Schmidt, 2009; Schneider, 2015; Stroebe & Strack, 2014; Zwaan et al., 2018). This is not to say that exact replications are impossible in other situations. In particular, exact replications may be possible and meaningful in situations in which the population and associated effect are assumed to be unchanging, and the sampling procedure, testing conditions, stimuli, measures, data aggregation approach, and analyses can all be duplicated from one test to the next while holding constant all other potentially influential factors. As I discuss further below, industrial quality control acceptance procedures provide a good example that meets these criteria.

Direct Replications are Possible in Social Science

Although *exact* replications are not possible in social science, *close* or *direct* replications are possible, and they are regarded as being essential for scientific progress (Brandt et al., 2014; Zwaan et al., 2018). Unlike exact replications, direct replications do not need to duplicate *all* aspects of the research methodology that might potentially influence the original effect. They only need to duplicate those aspects of the methodology that are currently regarded as being *theoretically essential* to reproduce the original effect (De Ruiter, 2018; Klein, 2014, p. 328; LeBel et al., 2017, p. 255; Nosek et al., 2012, p. 626; Zwaan et al., 2018, p. 4). These theoretically essential aspects are identified in "a theoretical commitment based on the current understanding of the phenomenon under study, reflecting current beliefs about what is needed to produce a finding" (Nosek & Errington, 2017, as cited in Zwaan et al., 2018, p. 3; see also LeBel et al., 2017, p. 255; Open Science Collaboration, 2015).

Hence, an exact replication needs to recreate all of the theoretically essential and potentially influential elements of the original study, whereas a direct replication needs to recreate

only the theoretically essential elements of the original study, and it can allow potentially influential but theoretically extraneous elements to vary from one study to the next. Indeed, it is important to allow potentially influential elements to vary between different studies in order to confirm their lack of influence and demonstrate the generality of the putative effect (Anderson et al., 2016; Lindsay & Ehrenberg, 1993, p. 220; Schmidt, 2009, p. 92). For example, a social scientist can undertake a series of direct replications in order to demonstrate that their proposed effect generalizes to different groups of participants who are tested at different times and in different places using different measures and different types of analyses. If any of these variables are considered theoretically essential for demonstrating the putative effect, then the researcher needs to hypothesize moderating effects that constrain the generality of the effect (De Ruiter, 2018; Rubin, 2017b, p. 315; Simons, 2014, p. 76; Simons, Shoda, & Lindsay, 2017). For example, in the case of historical changes in the gender difference in self-esteem, researchers need to develop theories that integrate sociological, cultural, and psychological variables in order to predict when and where the gender difference will and will not occur (e.g., Greenfield, 2017).

Different Types of Type I Error Probability

The replication crisis has generated a discussion about not only the feasibility of implementing exact and direct replications but also the importance of detecting Type I errors during hypothesis testing (for a review, see Shrout & Rodgers, 2018). This second discussion has focused on a particular type of Type I error that is based on the Neyman-Pearson approach to hypothesis testing (Neyman & Pearson, 1928, 1933; Nosek, Ebersole, DeHaven, & Mellor, 2018).[2] Below, I consider some interpretational difficulties with this Neyman-Pearson approach in the context of scientific disciplines that do not permit exact replications.

The Neyman-Pearson Type I Error Rate

Imagine a series of equally sized samples that all belong to one of two populations: a null population or an alternative population. Further imagine a researcher who does not know which population the samples belong to. The researcher's null hypothesis is that the samples belong to the null population, and their alternative hypothesis is that the samples belong to the alternative population. The researcher conducts a test in which they randomly draw one sample, measure the sample data, and compute a test statistic value and accompanying p value. In this scenario, the p value indicates the probability of obtaining a test statistic value that is as extreme or more extreme as the current value if exact replications of the test were to be reconducted using other samples in the series and assuming that the null hypothesis is correct. The researcher compares this p value to a prespecified significance threshold, or *alpha level*. If the p value falls at or below the alpha level (e.g., $p \leq .050$), then they declare their result to be "significant," and they decide to behave as if the series of samples belongs to the alternative population rather than the null population. In other words, they reject the null hypothesis and accept the alternative hypothesis. This approach is intended to limit, or *control*, the maximum frequency with which the test would lead to an incorrect decision to reject the null hypothesis if a long run of exact replications of the test were to be carried out on the other samples in the series. If the test's alpha level is set at .050 (i.e., reject the null hypothesis if $p \leq .050$), then it is assumed that the test's random measurement error (i.e., the random discrepancy between the measured sample and the true sample) and random sampling error (i.e., the random discrepancy between the true sample and the true population) would cause a sample from the null population to yield a test statistic value at least as extreme as the current value in no more than 5.00% of this long run of exact replications (Meehl, 1967, p. 104). Hence,

the alpha level controls the Type I error rate such that, "in the long run of experience, we shall not too often be wrong" (Neyman & Pearson, 1933, p. 291). Note that this approach does not allow us to know the probability of making a Type I error in relation to the particular sample that has been tested. It only indicates the maximum rate of making a Type I error across a long run of exact replications of a test that is reperformed on a series of samples (Neyman, 1971, p. 13; for a list of common misinterpretations of "significant results," see Gigerenzer, Krauss, & Vitouch, 2004, pp. 2-3).

Importantly, the Neyman-Pearson Type I error rate only applies to a long run of a test if the test's methodology remains fixed and unchanged and the sampling is random (Neyman, 1937, pp. 334-335; Neyman & Pearson, 1928, p. 177, p. 231, p. 232). Indeed, Neyman and Pearson (1928) stressed that "the limitation implied by the assumption of perfect random sampling must not of course be overlooked" (p. 232). When a test's methodology is fixed and the sampling is random, the only possible reasons for incorrectly rejecting the null hypothesis based on a test of a particular sample are random measurement error and random sampling error. However, if any potentially influential aspect of the testing methodology and/or sampling procedure changes from one test to the next, then the alpha level becomes inapplicable because the change may lead to samples being drawn from populations other than the specified, "admissible," null and alternative populations (Neyman & Pearson, 1933, p. 294). To illustrate, imagine that a researcher conducts a direct replication of their test in which they change a potentially influential part of the test (e.g., the sampling procedure, testing conditions, stimuli, or measures). In this case, the change may result in samples being drawn from a population other than the prespecified null and alternative populations. Hence, any incorrect decision to reject the null hypothesis and accept the alternative hypothesis may now be attributed to either (a) random measurement and sampling error causing a sample from the null population to appear like a sample from the alternative population or (b) a sample from some third, unspecified population appearing like a sample from the alternative population. It is for this reason that the Neyman-Pearson alpha level is only applicable to exact replications of the same test. It does not apply to direct replications, because direct replications may sample from different inadmissible populations.

It should be noted that it is possible to compute an *average* alpha level for a range of *different* tests that each refer to a different set of admissible populations (Neyman, 1977, pp. 108-109). For example, if the alpha levels of each of three different tests were .05, .05, and .05, or even .10, .05, and .001, then the average alpha level would be .05. However, this average alpha level does not necessarily apply to each of the specific tests. It merely indicates the mean alpha across the tests. Hence, it remains the case that each individual test's specific alpha level applies only to an exact replication of that individual test.

It also is possible to conceptualize the average alpha level as the Type I error rate for a combination of different tests that provide a test of a combined null hypothesis. However, in this case, the average alpha level would apply to a long run of exact replications of a combined test that draws samples from a single combined population.

Hence, whichever way the alpha level is conceptualized, it only applies to an exact replication of the associated test. Any deviation from the test's methodology opens up the possibility that the incorrect rejection of the null hypothesis is due not only to random measurement and sampling error but also to sampling from an inadmissible population that was not specified by the original test.

In summary, the Neyman-Pearson alpha level indicates the maximum frequency of making a Type I error if, and only if, a test was to be repeatedly reconducted on a long series of different

random samples that are all drawn from the exact same null population. This conceptualization of the Type I error probability is most appropriate in situations in which it is possible to undertake a long run of exact replications (Fisher, 1955, p. 69-70; see also Neyman, 1950, p. 331-335; Szucs & Ioannidis, 2017, p. 2). Such situations are typified by quality control tests in the context of industrial production (Fisher, 1955; Gigerenzer, 1993; Pearson, 1937, p. 54; see also Chow's, 1998, concept of utilitarian experiments). For example, a manufacturer might apply a quality control check on a factory's production line in order to test for a fault in a particular model of mobile phone (e.g., do the phones have a longer than expected start-up time?). In this situation, the manufacturer is able to design a replicable sampling process that ensures that every item in the objective, well-defined population (i.e., every phone on the production line) has a known probability of being sampled (e.g., Gigerenzer, 1993, p. 320; Neyman, 1950, p. 332; see also Berk et al., 1995, p. 432-433). And, for each test, the manufacturer is able to duplicate all of the factors that may potentially influence the test result (testing conditions, measure of start-up time, analyses, etc.). Consequently, it is meaningful to consider the long run Type I error rate in industrial quality control situations because, in these situations, the rejection of the null hypothesis may only be explained as being due to either a true positive (i.e., a production line of faulty mobile phones) or a false positive (i.e., a sample of non-faulty phones that are mischaracterised as being faulty). There is no opportunity to explain the rejection of the null hypothesis in terms of any other potentially influential factors (e.g., a change in the way in which the quality control check process is carried out that has resulted in a different model of phone being sampled from a different production line).

As previously established, exact replications are not possible in social science. In particular, social scientists are usually unable to design a random sampling process that ensures that every individual from the target population of interest (e.g., men and women) has a known probability of being sampled (Berk & Freedman, 2003; Berk et al., 1995; Gigerenzer & Marewski, 2015; Hacking, 1965, p. 125; Krause, 2019; Macdonald, 1997, p. 340; see also Greenland, 2006). Instead, they tend to use a convenience sampling approach and, following ethical guidelines, they only include participants who self-select to participate in the study and provide their informed consent. However, the Neyman-Pearson approach is not valid when "the sampling has not been random" (Neyman & Pearson, 1928, p. 177; see also Gigerenzer, 2004, p. 599; Hacking, 1965, p. 99-101; Ludbrook & Dudley, 1998, p. 127; Papineau, 1994, p. 443; Neyman, 1937, pp. 334-335; Neyman & Pearson, 1928, p. 232; Seidenfeld, 1979, p. 33; Shaver, 1993; Sterba, 2009; Strack, 2017).[3] Consequently, and unlike the quality control situation, it is not possible to control the long run Type I error rate in relation to the target population of interest (e.g., "men and women"; Frick, 1998; Greenland, 2006; Shaver, 1993). It is only possible to control this error rate in relation to the population that is actually randomly sampled. In social science, if random sampling occurs at all, then it occurs in relation to a very specific, transient, and potentially biased subpopulation of the target population of interest. For example, in order to investigate a proposed gender difference in self-esteem, a researcher might design a sampling procedure that randomly samples from the 2020 cohort of male and female undergraduate psychology students at an American university. In this case, the researcher is considering a time- and location-specific *statistical* null hypothesis in order to assess the more general *substantive* null hypothesis that men do not have greater self-esteem than women (e.g., Chow, 1998; Hager, 2013; Hurlbert & Lombardi, 2009, pp. 335-337; Meehl, 1967; Neyman, 1950). The associated significance test allows the researcher to determine whether the gender difference in self-esteem among 2020's male and female undergraduate psychology students at the American university is sufficiently large to decide to behave as if the sample was

not drawn from a corresponding statistical null population such that "in the long run of experience, we shall not too often be wrong" (Neyman & Pearson, 1933, p. 291). But the big question is why the researcher should be concerned about being wrong about a *long run* of tests of this particular statistical null hypothesis when the hypothesis has no potential to be retested in a series of exact replications. After all, the 2020 cohort of male and female undergraduate psychology students at the American university is a transient and constantly changing population (Rubin, 2017a; Schmidt, 2009; Serlin, 1987), and the Neyman-Pearson long run Type I error rate is not applicable when "the sampling has not been random or…*the population has changed during its course*" (Neyman & Pearson, 1928, p. 177, my emphasis). Furthermore, the psychological context associated with this particular statistical hypothesis is time- and location-specific and, consequently, not repeatable (Ferguson et al., 2014; Hansen, 2011; Schmidt, 2009; Van Bavel et al., 2016; Zuckerman et al., 2016). And yet it is only meaningful to control the Neyman-Pearson Type I error rate in relation to a testing methodology that has the capacity to be repeated across a long run of exact replications that hold all potentially influential factors constant, including the psychological context.

Of course, it is possible to *imagine* a hypothetical long run of exact replications of a specific statistical test. Indeed, social scientists can imagine a perfectly exact series of replications in which all potentially influential aspects of the methodology remain constant over an infinite number of tests in which samples are randomly drawn from a closed and unchanging population (Camilli, 1990, p. 138; Ludbrook & Dudley, 1998, p. 127). However, it is not useful or meaningful to imagine this scenario because it has no bearing on the reality of the social world or scientific practice (see also Berk & Freedman, 2003; Berk et al., 1995). As explained above, discussions following the replication crisis have concluded that many scientific disciplines proceed on the basis of direct replications rather than exact replications. Hence, researchers in these disciplines who use the Neyman-Pearson approach need to ask themselves why they are concerned about a long run Type I error rate that applies to a hypothetical series of exact replications when they are conducting a time- and context-specific test that refers to a transient and constantly changing population. In other words, researchers should consider why they are concerned about the long run Type I error rate for a "frequency of events in an endless series of repeated trials which will never take place" (Fisher, 1959, p. 101; see also Grayson, 1998).

Neyman and Scott (1958) discussed this issue when considering the application of the Neyman-Pearson statistical approach to cosmology:

> We are reminded that our Universe is unique. On the other hand, a statistical approach suggests questions and assertions formulated in terms of frequencies in repeated trials…if one contemplates the Universe as a single realization of a chance mechanism, it may appear impossible to subject any assertion about this chance mechanism to a test because it is impossible to repeat the experiment (p. 38).

Neyman and Scott's (1958) response to this problem was to argue that events in the universe are just as nonrepeatable as the sequence of outcomes in a roulette game, but we can treat groups of these events as being repeatable. For example, they argued that, "in order to treat toulette [sic] games indeterministically, we cut the total sequence in sections of convenient length and consider them as replications of the same experiment, with the same chance mechanism behind it" (Neyman & Scott, 1958, p. 39). However, this approach is only appropriate if it is plausible to assume that the same chance mechanism (cause or data-generating mechanism) that determines one section of the sequence of events also determines the other sections of the sequence. In the case of a game of roulette, this assumption is reasonable. In contrast, in the case of, for example, a game of poker, this assumption is problematic because it is understood that, as the game

proceeds, the players accumulate the history of the responses of the other players, and they learn and adapt their own responses as they go. Consequently, the chance mechanism that determines the outcomes in the first five hands of the poker game will be different to the chance mechanism that determines the outcomes in the last five hands. Hence, we return to the problem that, in some cases, "it is impossible to repeat the experiment" (Neyman & Scott, 1958, p. 38) because the chance mechanism has changed or, as Neyman and Pearson (1928) put it, "the population has changed during its course" (p. 177).

Again, the above points do not imply that the Neyman-Pearson approach is incorrect, only that it is more appropriate in some scenarios than in others (Hubbard, 2004, p. 300; Hubbard, 2011; Hurlbert & Lombardi, 2009; Gigerenzer & Marewski, 2015; Perezgonzalez, 2015a, p. 8; Perezgonzalez, 2017). Specifically, the Neyman-Pearson approach is most appropriate in situations in which it is possible to randomly sample from the entire target population of interest using the exact same testing procedure while holding potentially influential factors constant and when it makes sense to conceive the underlying chance mechanism as being immutable in "the long run of experience." In turn, these possibilities make it meaningful to control the long run Type I error rate. In contrast, in social science, (a) it is only possible to randomly sample from a transient and parochial subpopulation of the target population of interest, (b) it is not possible to repeat the testing procedure without changes in potentially influential factors, and (c) it does not make sense to consider the underlying chance mechanism (cause) as being fixed across time. Consequently, it is not meaningful to consider the long run Type I error rate. It is for this reason that Fisher characterised the Neyman-Pearson Type I error rate as "irrelevant" and "misleading" in scientific contexts (Fisher, 1926, p. 100; Fisher, 1955, p. 70; Fisher, 1958, p. 272; Fisher, 1959, p. 101; Fisher, 1961, p. 3; see also Perlman & Wu, 1999, pp. 364-365).

The Fisherian Type I Error Probability

Type I error probability is an important concept when researchers make decisions about hypotheses, and I am not suggesting that we abandon it. My concern is only with the way in which the concept is operationalized. In social science, it is not meaningful to operationalize the Type I error probability in relation to a series of samples that could have been randomly drawn from the exact same null population. Instead, it is more meaningful to consider the Type I error probability in relation to a single, time- and location-specific sample. Fisher's (1922, 1958, 1959) approach to significance testing allows a consideration of this sample-specific type of Type I error probability.

Although the Neyman-Pearson approach to hypothesis testing is often regarded as the more superior, modern replacement of the Fisherian approach, the Fisherian approach has recently enjoyed a revival (Haig, 2017, 2018; Hubbard, 2004, 2011; Hubbard & Bayarri, 2003; Hurlbert & Lombardi, 2009; Perezgonzalez, 2015a, 2017; Schneider, 2015). Like the Neyman-Pearson approach, the Fisherian approach compares an observed p value to a significance threshold (e.g., $p \leq .050$).[4] However, unlike Neyman-Pearson tests of *statistical* null hypotheses, the Fisherian approach treats p values that fall at or below this threshold as evidence against broader *substantive* null hypotheses.[5] Hence, if a researcher observes a p value (e.g., $p = .025$) that falls below their significance threshold (e.g., $p \leq .050$), and they assume, counterfactually, that all necessary statistical and methodological assumptions have been met, then they can accept the p value as providing a preliminary piece of evidence against the substantive null hypothesis, and they can adopt the "provisional" (i.e., initial but changeable) attitude that the substantive null hypothesis should be discounted (Fisher, 1955, p. 74,; Fisher, 1959, p. 42, p. 100; Hubbard, 2011; Macdonald, 1997, p. 339).[6] The smaller the p value relative to the significance threshold, the greater the

preliminary evidence against the substantive null hypothesis, and so the stronger the provisional attitude that this hypothesis should be discounted.

Importantly, Fisher was firmly against the Neyman-Pearson concept of a long run Type I error rate (e.g., Fisher, 1959, p. 42, p. 100) because it relies on the assumption of "repeated sampling from the same population" (Fisher, 1955, p. 71; Fisher, 1959, p. 78, p. 83) and, given the reference class problem (Venn, 1876), it is often difficult for the researcher to know "which population is to be used to define the probability level" (Fisher, 1959, p. 71; Gigerenzer, 1993, p. 320). For example, when investigating gender differences in self-esteem among 2020's male and female undergraduate psychology students at an American university, the null population may refer to "men and women," "twenty-first century men and women," "young men and women," "educated men and women," "American men and women," and so on. In cases such as this, Fisher argued that researchers need to imagine a reference class (e.g., "men and women") on the understanding that they may be wrong and that some other reference class is correct (e.g., "young men and women"). He argued that this reference class uncertainty makes the concept of a long run Type I error rate irrelevant because it may refer to an incorrect reference class. Instead of considering the long run Type I error rate during repeated sampling from the same population, researchers should conceive each new sample as coming from a potentially different population (e.g., young vs. old men and women). Hence, researchers should always ask themselves: "of what population is this a random sample?" (Fisher, 1922, p. 313).

Despite his strong opposition to the Neyman-Pearson Type I error rate, Fisher was in favour of researchers considering the probability that they had made an error in provisionally rejecting a null hypothesis (e.g., Fisher, 1937, p. 16; Fisher, 1959, p. 35; see also Fisher, 1959, pp. 100-101). Indeed, and as others have pointed out, the Fisherian significance threshold provides a basis for computing a Type I error *probability*, but one that is a conceptually different to the type of Type I error *rate* provided by the Neyman-Pearson approach (Macdonald, 1997, p. 339; Mayo, 2014; Perezgonzalez, 2015a, p. 5, Perezgonzalez, 2017, p. 8; Royall, 1997, p. 86). In particular, the Fisherian Type I error is a *sample-specific* probability in that it is conditioned on a hypothetical population that is imagined to reflect the characteristics of the particular sample under investigation rather than on a series of random samples from a well-defined objective population. In the Fisherian approach, researchers make provisional (preliminary) decisions about rejecting substantive null hypotheses based on a specific sample of data, and these provisional decisions are guided by whether observed *p*-values fall below researchers' significance thresholds (e.g., Fisher, 1937, p. 16; Fisher, 1955, p. 74; Fisher, 1959, p. 35, p. 100-101). Consequently, Fisherian researchers can also make provisional Type I errors in relation to their decisions (Cox & Hinkley, 1974, p. 66; Royall, 1997, p. 73) and, assuming that all necessary statistical and methodological assumptions are met, the significance threshold indicates the probability of making this error with regards to a specific sample of data.[7] If an observed *p* value falls below a researcher's significance threshold of $p \leq .05$, then the researcher may decide to provisionally reject the substantive null hypothesis on the understanding that they have a 5.00% probability of making an error due to "an exceptionally rare chance" that has occurred in relation to the particular sample of data under consideration (Fisher, 1959, p. 39).

Again, Fisherian sample-specific Type I error probabilities are quite different from Neyman-Pearson long run Type I error rates because they do not refer to a maximum frequency of incorrectly rejecting a statistical null hypothesis across a long run of exact replications (for related discussions, see Berger & Delampady, 1987, p. 329; Hubbard, 2004, 2011; Hubbard & Bayarri, 2003, p. 174; Fisher, 1955, p. 71-72; Fisher, 1959, p. 78; see also Fisher, 1962, p. 530; Goodman,

1999, p. 999; Heike, Târcolea, Tarcolea, & Demetrescu, 2004, p. 5; Hubbard, 2004, 2011; Johnstone, 1987, p. 483; Neyman, 1971, p. 13; Royall, 1997, pp. 41-50). Instead, a Fisherian Type I error probability is conditioned on "a population of samples in all relevant respects like that observed [excluding the test results]" (Fisher, 1955, p. 72). Critically, this imaginary *hypothetical infinite population* (Fisher, 1922, p. 311) *does not contain any recognizable relevant subsets to which different error probabilities may apply*. The concern about recognizable relevant subsets is a central but often overlooked aspect of Fisher's approach (Camilli, 1990, p. 137; Johnstone, 1987, p. 485; Johnstone, 1989; Seidenfeld, 1979; Senn, 2005), and it is integral to his notion of *fiducial probability* and his *fiducial argument* (Pedersen, 1978, p. 152).[8] In order to apply a Fisherian significance threshold (e.g., $p \leq .050$) in relation to a population, researchers need to assume ("imagine") that the population does not include any theoretically relevant subpopulations (e.g., born in the U.S.A. vs. born in China; born before 2000 vs. born after 2000, high or low on sexism, etc.) that could give rise to substantively different probability statements. The variables that demarcate these relevant subsets (e.g., culture; age) might be described as *hidden moderators* (Zwaan et al., 2018). It is this "postulate of ignorance" about relevant subsets in the population (Fisher, 1958, p. 268; Fisher, 1959, pp. 32-33, p. 57) that allows the move from a Type I error rate across the long run (an aleatory, frequentist form of probability) to a Type I error probability in the current case (an epistemic, sample-specific form of probability; Johnstone, 1987, 1989). Specifically, the postulate of ignorance allows researchers to legitimately (logically) attach the significance threshold and its associated Type I error probability to samples that are drawn from an imaginary (hypothetical), sample-specific population, or *reference set*, rather than to samples that are drawn during an imaginary long run of exact replications from the exact same, fixed, well-defined, objective population (delineating the frequentist *sample space*). As Fisher (1959, p. 83) explained, the observed sample "is not one of an objective series of similar samples from the same population existing in reality, though it can be regarded by an act of imagination as one of a hypothetical reference set." This hypothetical reference set is conditioned on an *ancillary statistic* that has the same value as that of the observed sample (Cox, 1958, pp. 359-361; Fisher, 1955, pp. 71-72; Johnstone, 1987, p. 482; Lehmann, 1993, p. 1245-1246; Pedersen, 1978, p. 152). So, for example, looking at Fig. 1, the Fisherian significance threshold (e.g., $p \leq .050$) may indicate the probability of making an incorrect provisional decision to reject the substantive null hypothesis that men do not have greater self-esteem than women based on the unique situation in which self-esteem is measured among a sample of "people" (reference set) from a parochial population of first-year, undergraduate, psychology students at the University of X and under the assumption that this reference set does not contain any recognizable relevant subsets. As Fisher (1959, p. 23) explained, "[relevant] subsets must always exist; it is required that no one of them shall be recognizable." Hence, in Fig. 1, although Sample A may be drawn from a relevant subset (e.g., "psychology students," "undergraduate students," "people born during 1990-1995," or "people high in sexism"), the researcher must not be able to recognize these subsets as being relevant to their Type I error probability statement. In other words, the researcher should have no theoretical or empirical grounds for suspecting that these factors make a difference.

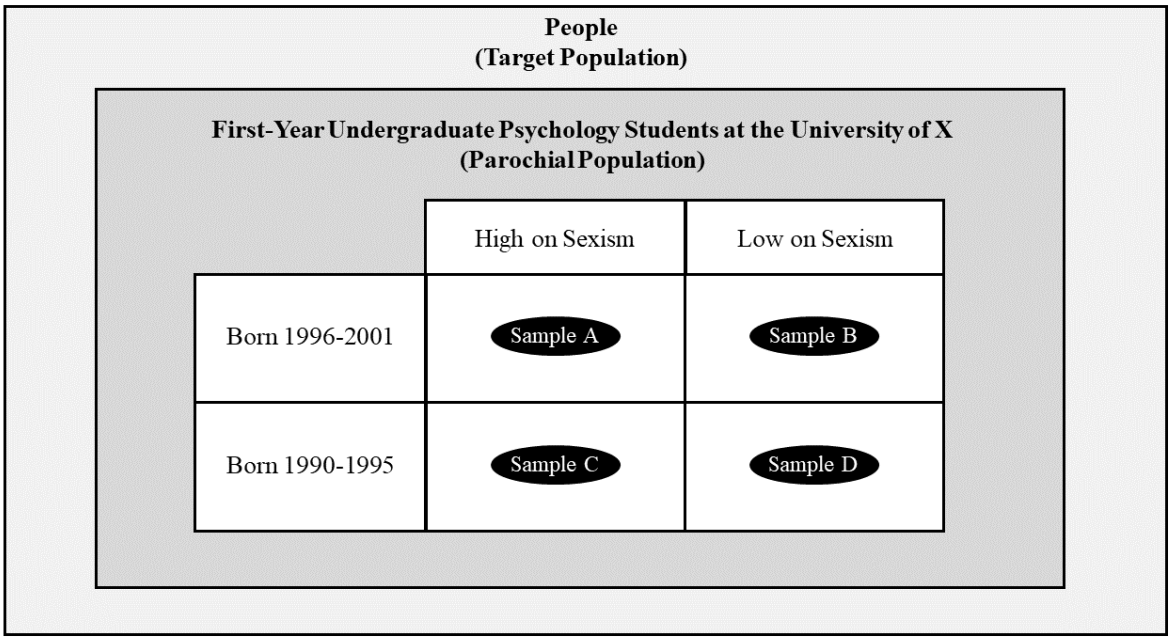

Fig. 1. Imagine that a researcher tests for a gender difference in self-esteem. In this case, a Fisherian researcher would imagine a reference set that does not contain any potentially relevant subsets. For example, they might imagine that year of birth (1990-1995 or 1996-2001), sexist attitudes (high or low), degree subject (psychology), and university (the University of X) do not affect the size of the gender difference in self-esteem. In this scenario, Samples A to D are considered to be part of the same hypothetical infinite population of "people." In contrast, a Neyman-Pearson researcher would conceive their test as having the potential to repeatedly sample from the same, well-defined, objective, parochial population, which in this case is "first-year undergraduate psychology students at the University of X." A Neyman-Pearson researcher would also assume that their test is able to sample from all potentially influential (relevant) subsets within that population (i.e., Samples A, B, C, and D).

Again, a key implication of the Fisherian approach is that Type I error probabilities are sample-specific (Ludbrook & Dudley, 1998, pp. 128-129). If a researcher actually drew another sample of participants from the parochial population, then they might unwittingly draw it from a relevant subset of that population that has a different ancillary value and that does not conform to the original Type I error probability. For example, in Fig. 1, a replication conducted five years after the original study might be more likely to draw Samples A and B rather than Samples C and D. Hence, the significance threshold and its associated Type I error probability always need to be interpreted in relation to an imagined null population that is predicated on the characteristics of the observed sample and that does not contain any recognisable relevant subsets. The sample can then be conceived as being a random sample from this hypothetical infinite null population (reference set; Fisher, 1922, p. 311; Fisher, 1958, pp. 263-264; Gigerenzer, 2006, p. 245; Johnstone, 1987, p. 497; Johnstone, 1989; Sterba, 2009, p. 716).

It might be argued that there is something circular in conditioning probability statements on hypothetical infinite populations that are imagined to resemble observed samples. However, this act of imagination merely serves to alert researchers to the fact that their results do not necessarily generalize to other populations, as represented by other samples.

It should also be noted that the Neyman-Pearson and Fisherian approaches both require an act of imagination on the part of the researcher. However, the implications of these acts of imagination are quite different. In the Neyman-Pearson case, the act of imagination is to believe

that it is possible to repeatedly sample from the exact same parochial population (e.g., first-year undergraduate psychology students at the University of X). As discussed above, this belief is theoretically unreasonable and empirically unrealistic in the case of people, social groups, and social systems because these units of investigation change over time (e.g., the psychology department at University of X may expand to a new campus in a new country). In contrast, in the Fisherian case, the act of imagination is to assume that the hypothetical reference set does not contain any relevant subsets that would give rise to alternative probability statements. In the absence of any theory or evidence that would allow the recognition of such relevant subsets, this assumption is reasonable and scientifically useful (Senn, 2005). Furthermore, as explained below, the validity of this assumption can be tested in a series of real future studies that sample from similar populations (direct replications) and different populations (conceptual replications) but never from the same population (exact replications).

Contrasting the Neyman-Pearson and Fisherian Approaches

In summary, the Neyman-Pearson Type I error rate refers to a replicable random sampling procedure that has the potential to sample from all relevant subsets in the target population of interest during a long run of exact replications. In contrast, the Fisherian Type I error probability refers to a hypothetical sampling procedure that is restricted to an imaginary sample-specific population (reference set) that does not contain any recognizable relevant subsets. Which of these error probabilities is more appropriate in social science? I argue that there are four reasons that the Fisherian sample-specific Type I error probability is more appropriate than the Neyman-Pearson long run Type I error rate.

First, and as discussed above, social scientists investigate irreversible units in the form of people, social groups, and social systems. Consequently, it is impossible for social scientists to conduct exact replications because potentially influential factors will always vary from one study to the next. In this context, it is more realistic to consider a Fisherian Type I error probability for the specific sample under investigation than it is to consider a Neyman-Pearson Type I error rate for a long run of exact replications.

Second, and as a result of the first reason, social scientists should condition their Type I error probabilities on the sample that they actually observed rather than on samples that they could have observed in a long run of replications. The Neyman-Pearson long run Type I error rate does not meet this *conditionality principle* (Cox, 1958, p. 359-361; Lehmann, 1993, p. 1245-1246; Wagenmakers, 2007, p. 783). Instead, the Neyman-Pearson approach provides an *unconditional* Type I error rate that applies across all of the potentially relevant subsets within a population (e.g., in Fig. 1, across people born during 1990-1995 and 1996-2001 as well as across people who are high and low on sexism). This unconditional approach is problematic because the current sample may be drawn from a subset of the population that is substantively different to the other subsets in the population. To address this problem, the Fisherian Type I error probability is conditioned on the current sample and only applies to a corresponding hypothetical infinite population (reference set) that does not contain any subsets that the researcher recognizes to be relevant to the statistical inference in question. Following this conditional approach, the relevance of population subsets may become recognized (discovered) during the course of a series of direct or conceptual replications. For example, sexist attitudes may be recognized as an important moderator of the gender difference in self-esteem such that the gender difference is stronger among people who are high in sexism.

Third, social scientists are not able to randomly sample the entire target population of interest (e.g., men and women). Instead, they are only able to sample from parochial populations (e.g., 2020's male and female undergraduate students at an American university). Consequently, they are not able to undertake a comprehensive assessment of potential moderating variables in a single study (i.e., variables that demarcate relevant subsets in the population). Instead, social scientists need to conduct a series of actual direct replications across a range of parochial populations that all share the defining characteristics of the broader target population of interest (e.g., men and women in the U.S.A, men and women in China, men and women in France, etc.; Hurlbert, & Lombardi, 2009, pp. 336-337; Fisher, 1937, p. 16). They then need to collate their results in order to reach firmer conclusions about the conditions under which the broader substantive null hypothesis is and is not rejected (e.g., by conducting a meta-analysis that includes a moderation analysis). The Fisherian Type I error probability is better suited to this piecemeal cumulative approach to knowledge building because it can be applied to each sample-specific, provisional, belief-changing decision that researchers make about rejecting the same broad substantive null hypothesis in a series of direct and conceptual replications. In contrast, the Neyman-Pearson long run Type I error rate is less useful in this context because it refers to an unconditional, final decision about rejecting a parochial, context-specific, statistical null hypothesis that is specified across a long run of exact replications.

Finally, social scientists need to follow Fisher's postulate of ignorance when they make a "theoretical commitment" that their observed results will replicate in direct replications (LeBel et al., 2017, p. 255; Nosek et al., 2012, p. 626; Zwaan et al., 2018, p. 4). In particular, a researcher who specifies the elements of their original study that are essential for a direct replication also needs to concede a "subjective ignorance" (Fisher, 1959, p. 33) about potentially influential but theoretically extraneous elements that may vary from one study to the next (e.g., whether participants' culture or age will have any effect on the results; for discussions, see Cesario, 2014; Earp & Trafimow, 2015, p. 3). If the researcher is not ignorant about a particular element, and they are able to recognize it as a theoretically relevant subset (e.g., the effect should only hold for heterosexual students), then they need to make that subset their reference set and declare it as an essential element for any subsequent direct replication (Fisher, 1959, p. 111; Johnstone, 1987, 1989; Senn, 2005; for similar reasoning, see De Ruiter, 2018; Rubin, 2017b, p. 315; Simons, 2014, p. 76; Simons, Shoda, & Lindsay, 2017). Otherwise, they must claim a subjective ignorance about potentially theoretically relevant subsets and associated hidden moderators.

Consistent with Fisher's emphasis on subjective ignorance, the scientific literature is replete with cases in which researchers have discovered effects that were initially assumed to be quite general and then found to be qualified by previously hidden moderators such that their generality in size or existence became more circumscribed as research progressed (for a discussion and examples, see Firestein, 14/02/2016; Redish, Kummerfeld, Morris, & Love, 2018). As Redish et al. (2018, p. 5043) explained, "in many…cases, what have been called 'failures to replicate' are actually failures to generalize across what researchers hoped were inconsequential changes in background assumptions or experimental conditions." Similarly, Nosek and Lakens (2014, p. 138) explained that "different results between original and replication research could mean that there are unknown moderators or boundary conditions that differentiate the two studies" (see also Open Science Collaboration, 2015, p. 6; see also Camerer et al., 2018; Zwaan et al., 2018, p. 4). To be clear, the discovery of these hidden moderators (relevant subsets) *does* provisionally falsify (lower the estimated relative verisimilitude of) the original unconditional hypothesis. Nonetheless, the acknowledgment of such previously hidden moderators allows the development of a new

hypothesis. In some cases, this new hypothesis may represent a progressive update in the theoretical explanation for the originally hypothesised effect (e.g., moderator variable X is introduced into the theory in order to predict when the originally hypothesised effect will and will not occur; Lakatos, 1976; McGuire, 1983, pp. 7-8; Zwaan et al., 2018, p. 4; see also the commentaries to Camerer et al., 2018). In some cases, the new hypothesis may refer to the methodology that is essential to produce the originally hypothesised effect (e.g., maintaining measurement error below a certain level; Duncan & Davachi, 2018). Finally, in some cases, the new hypothesis may entail a completely different explanation for the effect and sometimes one that is theoretically less interesting than the original hypothesis (e.g., explaining the effect in terms of methodological artefacts such as demand characteristics or stimulus sampling failures). None of these new hypotheses imply that the original effect was a Type I error. Instead, they represent alterative explanations for a genuine (true positive) effect.

It is important to appreciate that, in the Fisherian approach, the consideration of hidden moderators (relevant subsets) occurs *in addition to*, rather than instead of, the consideration of Type I and II errors.[9] Hence, assuming that all necessary statistical and methodological assumptions are met, a failure to replicate an effect in a direct replication may be due to (a) "a very remarkable and exceptional coincidence" in the initial study (i.e., a sample-specific Type I error; Fisher, 1959, p. 35), (b) a lack of sensitivity to detect the effect in the direct replication (i.e., a sample-specific Type II error; Fisher, 1937, p. 25), or (c) the operation of a previously unknown moderator variable that becomes recognised as demarcating a relevant subset in the population (Camerer et al., 2018; Open Science Collaboration, 2015, p. 6; Nosek & Lakens, 2014, p. 138; Nosek & Errington, 2019, p. 4; Zwaan et al., 2018, p. 4).

In summary, the Fisherian Type I error probability provides a means of evaluating each sample-specific result in a series of actual direct replications without reference to the results of a series of impossible and unobserved exact replications. Scientists who sample irreversible units from parochial populations, who condition their inferences on observed data per se, and who concede subjective ignorance about the essential elements of direct replications should find the Fisherian sample-specific Type I error probability more applicable to their research than the Neyman-Pearson long run Type I error rate.

Return to the Replication Crisis

Researchers have advocated two main approaches in response to the replication crisis. The first approach has been to improve the identification of Type I errors. The second approach has been to undertake direct replications. None of the arguments that are presented in this article oppose either of these approaches. Instead, the present article questions how social scientists should conceptualize Type I errors given that exact replications are impossible in social science and direct replications are possible and essential.

I considered two types of Type I error probability. The Neyman-Pearson alpha level limits the maximum frequency of Type I errors that would occur in a long run of exact replications of a test that was reconducted on a series of samples from the same objective population. In contrast, the Fisherian significance threshold indicates the probability of making a Type I error about the provisional decision to reject a substantive null hypothesis on the basis of a specific piece of evidence from a specific sample of a hypothetical population. Of these two approaches, the Neyman-Pearson approach is least applicable in social science because social science deals with irreversible units (viz., people, social groups, and social systems) that make exact replications impossible and, therefore, long run error rates meaningless.

The Fisherian sample-specific Type I error probability is more appropriate in social science because it limits a consideration of the Type I error probability to a provisional decision about rejecting a substantive null hypothesis based on a single sample rather than to a final decision about rejecting a statistical null hypothesis in relation to a series of samples. In the Fisherian approach, assessments of replicability depend on the results of a series of real direct replications, and significance thresholds are used to assist researchers in making provisional decisions in each study. Putative effects can be investigated using this gradual, piecemeal, cumulative approach in order to (a) test their generality and (b) generate new hypotheses about their limiting conditions that may be included in revised theories (Firestein, 14/02/2016; Lakatos, 1976; McGuire, 1983, pp. 7-8, p. 14; Redish et al., 2018). In other words, the Fisherian approach "afford[s] direct guidance as to what elements we may reasonably incorporate in any theories we may be attempting to form in explanation of objectively observable phenomena" (Fisher, 1959, p. 35). Fisher described this process of progressive theoretical development as "learning by observational experience" (Fisher, 1937, p. 9; Fisher, 1955, p. 73; Fisher, 1959, p. 100-101; for a similar view, see McGuire's, 1983, contextualist vision of science). He argued that the Neyman-Pearson approach is not well-suited to learning by observational experience because its concept of long run error rates implies a fixed and closed system of "repeated sampling from the same population" (Fisher, 1955, p. 71; Fisher, 1959, p. 78, p. 83) in which "nothing essentially new can be discovered" (Fisher, 1959, p. 109; see also Cox, 1958, p. 360). In particular, the Neyman-Pearson long run Type I error rate is not suitable if one's aim is to discover (recognise) relevant subsets (hidden moderators) within the population.

Fisher predicted that "the principles of Neyman and Pearson's 'Theory of Testing Hypotheses' are liable to mislead those who follow them into much wasted effort and disappointment" (Fisher, 1959, p. 89). Indeed, it is possible to attribute part (but not all) of the replication crisis to researchers' unrealistic expectations about replication rates that are based on Neyman-Pearson long run error rates. For example, one of the ways in which the Open Science Collaboration (OSC, 2015, p. 4) computed a replication rate was to refer to the Neyman-Pearson concept of power:

> On the basis of only the average replication power of the 97 original, significant effects [M = 0.92, median (Mdn) = 0.95], we would expect approximately 89 positive results in the replications if all original effects were true and accurately estimated; however, there were just 35.

Because it is based on the Neyman-Pearson Type II error rate (i.e., 1 – power), this calculation is only valid for exact replications. However, the OSC studies were direct replications, not exact replications. As Gilbert, King, Pettigrew, and Wilson (2016, p. 1) explained, using the concepts of power and the Type II error rate in this way assumes

> that the one and only way in which OSC's replication studies differed from the original studies is that they drew new samples from the original population. In fact, many of OSC's replication studies differed from the original studies in other ways as well.

Indeed, Gilbert et al. (2016) reviewed the OSC studies and highlighted cases in which different populations and procedures were used compared to those used in the original studies. Hence, the OSC studies represented direct replications, rather than exact replications. Consequently, it is inappropriate to use Neyman-Pearson long run error rates to compute a replication rate for the OSC studies, and doing so may help to explain the "disappointment" that many researchers felt about the OSC replication rate (Fisher, 1959, p. 89).

To end, I want to make it clear that I am not dismissing the social reality of the replication crisis: There is quite obviously a crisis of confidence about the standard approach to science. However, I do think that it remains unclear as to what extent the crisis is due to scientifically problematic levels of replicability rather than researchers' unrealistic expectations about replicability that are based, in part, on a consideration of Neyman-Pearson long run error rates. In this respect, adopting the Fisherian sample-specific Type I error probability may be beneficial not only because the Fisherian approach is more consistent with the subject matter and research practices of social scientists, but also because it may help to reduce the emphasis on ultimately unachievable exact replication rates and instead increase the focus on the degree of evidence that has been obtained for and against a prospective effect in a series of unique studies.

References

Amrhein, V., Korner-Nievergelt, F., & Roth, T. (2017). The earth is flat (p> 0.05): Significance thresholds and the crisis of unreplicable research. *PeerJ, 5,* e3544. https://doi.org/10.7717/peerj.3544

Amrhein, V., Greenland, S., & McShane, B. (2019). Retire statistical significance. *Nature, 567,* 305-307.

Anderson, C. J., Bahník, Š., Barnett-Cowan, M., Bosco, F. A., Chandler, J., Chartier, C. R.,... Zuni, K. (2016). Response to "Comment on Estimating the reproducibility of psychological science". *Science, 351,* 1037-1039. https://doi.org/10.1126/science.aad9163

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E. J., Berk, R.,...& Cesarini, D. (2018). Redefine statistical significance. *Nature Human Behaviour, 2*, 6-10. https://doi.org/10.1038/s41562-017-0189-z

Berger, J. O. (2003). Could Fisher, Jeffreys and Neyman have agreed on testing? *Statistical Science, 18,* 1-32. https://doi.org/10.1214/ss/1056397485

Berger, J. O., & Delampady, M. (1987). Testing precise hypotheses. *Statistical Science*, 2, 317-335.

Berk, R. A., & Freedman, D. A. (2003). Statistical assumptions as empirical commitments. In T. G. Blomberg & S. Cohen (Eds.), *Law, punishment, and social control: Essays in honor of Sheldon Messinger* (2nd ed., pp. 235–254). New York: Aldine.

Berk, R. A., Western, B., & Weiss, R. E. (1995). Statistical inference for apparent populations. *Sociological Methodology, 25,* 421-458. https://doi.org/10.2307/271073

Billig, M. (2018). Those who only know of social psychology know not social psychology: A tribute to Gustav Jahoda's historical approach. *Culture & Psychology*, *24*, 282-293. https://doi.org/10.1177/1354067X18779042

Bowater, R. J. (2017). A defence of subjective fiducial inference. *AStA Advances in Statistical Analysis, 101,* 177-197. https://doi.org/10.1007/s10182-016-0285-9

Brandt, M. J., IJzerman, H., Dijksterhuis, A., Farach, F. J., Geller, J., Giner-Sorolla, R.,...& Van't Veer, A. (2014). The replication recipe: What makes for a convincing replication? *Journal of Experimental Social Psychology*, *50*, 217-224. https://doi.org/10.1016/j.jesp.2013.10.005

Brunswik, E. (1955). Representative design and probabilistic theory in a functional psychology. *Psychological Review, 62,* 193–217. https://doi.org/10.1037/h0047470

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T. –H., Huber, J., Johannesson, M.,…Wu, H. (2018). Evaluating the replicability of social science experiments in *Nature* and *Science*

between 2010 and 2015. *Nature Human Behaviour, 2,* 637-644. https://doi.org/10.1038/s41562-018-0399-z

Camilli, G. (1990). The test of homogeneity for 2× 2 contingency tables: A review of and some personal opinions on the controversy. *Psychological Bulletin, 108,* 135-145.

Cesario, J. (2014). Priming, replication, and the hardest science. *Perspectives on Psychological Science, 9,* 40-48. https://doi.org/10.1177/1745691613513470

Chow, S. L. (1998). Précis of statistical significance: Rationale, validity, and utility. *Behavioral and Brain Sciences, 21,* 169-194. https://doi.org/10.1017/S0140525X98261164

Cox, D. R. (1958). Some problems connected with statistical inference. *The Annals of Mathematical Statistics, 29,* 357-372. https://doi.org/10.1214/aoms/1177706618

Cox, D. R., & Hinkley, D. V. (1974). *Theoretical statistics*. London: Chapman & Hall.

Cumming, G. (2008). Replication and *p* intervals: *p* values predict the future only vaguely, but confidence intervals do much better. *Perspectives on Psychological Science, 3,* 286-300. https://doi.org/10.1111/j.1745-6924.2008.00079.x

De Boeck, P., & Jeon, M. (2018). Perceived crisis and reforms: Issues, explanations, and remedies. *Psychological Bulletin, 144,* 757-777. https://doi.org/10.1037/bul0000154

de Ruiter, J. (2018). *The meaning of a claim is its reproducibility. Behavioral and Brain Sciences, e125*. https://doi.org/10.1017/S0140525X18000602

Duncan, K., & Davachi, L. (2018). Disengagement with cognitive tasks decreases effect sizes. *Nature Human Behavior, 2,* 606. https://doi.org/10.1038/s41562-018-0409-1

Earp, B. D., & Trafimow, D. (2015). Replication, falsification, and the crisis of confidence in social psychology. *Frontiers in Psychology, 6,* 621. https://doi.org/10.3389/fpsyg.2015.00621

Ferguson, M. J., Carter, T. J., & Hassin, R. R. (2014). Commentary on the attempt to replicate the effect of the American flag on increased Republican attitudes. *Social Psychology, 45,* 299-311.

Fienberg, S. E., & Tanur, J. M. (1996). Reconsidering the fundamental contributions of Fisher and Neyman on experimentation and sampling. *International Statistical Review/Revue Internationale de Statistique, 64,* 237-253.

Firestein, S. (14/02/2016). Why failure to replicate findings can actually be good for science. *LA Times*. Retrieved from http://www.latimes.com/opinion/op-ed/la-oe-0214-firestein-science-replication-failure-20160214-story.html?outputType=amp&__twitter_impression=true

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 222,* 309-368. https://doi.org/10.1098/rsta.1922.0009

Fisher, R. A. (1926). *Statistical methods for research workers.* Edinburgh, UK: Oliver and Boyd.

Fisher, R. A. (1937). *The design of experiments* (2nd ed.). Edinburgh, UK: Oliver & Boyd.

Fisher, R. A. (1955). Statistical methods and scientific induction. *Journal of the Royal Statistical Society. Series B (Methodological), 17,* 69-78.

Fisher, R. A. (1958). The nature of probability. *The Centennial Review, 2,* 261-274.

Fisher, R. A. (1959). *Statistical methods and scientific inference* (2nd ed.). Edinburgh: Oliver & Boyd.

Fisher, R. A. (1961). Sampling the reference set. *Sankhyā: The Indian Journal of Statistics, Series A, 23,* 3-8.

Fisher, R. A. (1962). The place of the design of experiments in the logic of scientific inference. *Colloques Internationaux. Centre National de la Recherche Scientifique, Paris, 110*, 13-19. Retrieved from https://publications.csiro.au/rpr/pub?list=BRO&pid=procite:4a6b965a-3666-4914-92fc-786ec3983d60

Fraser, D. A. S. (2008). Fiducial inference. *International Encyclopedia of the Social Sciences.* Retrieved from https://www.encyclopedia.com/social-sciences/applied-and-social-sciences-magazines/fiducial-inference

Frick, R. W. (1998). Interpreting statistical testing: Process and propensity, not population and random sampling. *Behavior Research Methods, Instruments, & Computers*, 30, 527-535. https://doi.org/10.3758/bf03200686

Gigerenzer, G. (1993). The superego, the ego, and the id in statistical reasoning. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 311-339). Hillsdale, NJ: Lawrence Erlbaum.

Gigerenzer, G. (2004). Mindless statistics. *Journal of Socio-Economics, 33*, 587-606. https://doi.org/10.1016/j.socec.2004.09.033

Gigerenzer, G. (2006). What's in a sample? A manual for building cognitive theories. In K. Fiedler & P. Juslin, (Eds.), *Information sampling and adaptive cognition* (pp. 239-260). New York: Cambridge University Press.

Gigerenzer, G., Krauss, S., & Vitouch, O. (2004). The null ritual: What you always wanted to know about null hypothesis testing but were afraid to ask. In Kaplan, D. (Ed.), *The Sage handbook of quantitative methodology for the social sciences* (pp. 391-408). New York: Sage.

Gigerenzer, G., & Marewski, J. N. (2015). Surrogate science: The idol of a universal method for scientific inference. *Journal of Management, 41*, 421-440. https://doi.org/10.1177/0149206314547522

Gilbert, D. T., King, G., Pettigrew, S., & Wilson, T. D. (2016). Comment on "Estimating the reproducibility of psychological science". *Science, 351* (6277), 1037-1037. https://doi.org/10.1126/science.aad7243

Goodman, S. N. (1999). Toward evidence-based medical statistics. 1: The *P* value fallacy. *Annals of Internal Medicine, 130*, 995-1004. https://doi.org/10.7326/0003-4819-130-12-199906150-00008

Grayson, D. A. (1998). The frequentist façade and the flight from evidential inference. *British journal of Psychology, 89*, 325-345. https://doi.org/10.1111/j.2044-8295.1998.tb02687.x

Greenfield, P. M. (2017). Cultural change over time: Why replicability should not be the gold standard in psychological science. *Perspectives on Psychological Science, 12*, 762-771. https://doi.org/10.1177/1745691617707314

Greenland, S. (2006). Bayesian perspectives for epidemiological research: I. Foundations and basic methods. *International Journal of Epidemiology, 35*, 765-775. https://doi.org/10.1093/ije/dyi312

Greenland, S., & Chow, Z. R. (2019). *To aid statistical inference, emphasize unconditional descriptions of statistics*. arXiv preprint arXiv:1909.08583.

Hacking, I. (1965). *Logic of statistical inference*. London: Cambridge University Press.

Hager, W. (2013). The statistical theories of Fisher and of Neyman and Pearson: A methodological perspective. *Theory & Psychology, 23*, 251-270. https://doi.org/10.1177/0959354312465483

Haig, B. D. (2017). Tests of statistical significance made sound. *Educational and Psychological Measurement, 77*, 489-506. https://doi.org/10.1177/0013164416667981

Haig, B. D. (2018). *The philosophy of quantitative methods: Understanding statistics.* New York: Oxford University Press.

Hampel, F. R. (2003). *The proper fiducial argument.* Seminar für Statistik, Eidgenössische Technische Hochschule (ETH), Zürich. Retrieved from https://doi.org/10.3929/ethz-a-004526011

Hannig, J., Iyer, H., Lai, R. C., & Lee, T. C. (2016). Generalized fiducial inference: A review and new results. *Journal of the American Statistical Association, 111,* 1346-1361. https://doi.org/10.1080/01621459.2016.1165102

Hansen, W. B. (2011). Was Herodotus correct? *Prevention Science, 12,* 118-120. https://doi.org/10.1007/s11121-011-0218-5

Heike, H., Târcolea, C. T., Tarcolea, A. I., & Demetrescu, M. (2004). Fiducial inference: An approach based on bootstrap techniques. *Mimeo*. Retrieved from https://www.researchgate.net/profile/Matei_Demetrescu/publication/252660273_Fiducial_Inference_An_approach_based_on_bootstrap_techniques/links/541ffe7a0cf241a65a1af205.pdf

Hubbard, R. (2004). Alphabet soup: Blurring the distinctions between *p*'s and α's in psychological research. *Theory & Psychology, 14,* 295-327. https://doi.org/10.1177/0959354304043638

Hubbard, R. (2011). The widespread misinterpretation of *p*-values as error probabilities. *Journal of Applied Statistics, 38,* 2617-2626. https://doi.org/10.1080/02664763.2011.567245

Hubbard, R., & Bayarri, M. J. (2003). Confusion over measures of evidence (*p*'s) versus errors (α's) in classical statistical testing. *The American Statistician, 57,* 171-178. https://doi.org/10.1198/0003130031856

Hurlbert, S. H., & Lombardi, C. M. (2009). Final collapse of the Neyman-Pearson decision theoretic framework and rise of the neoFisherian. *Annales Zoologici Fennici, 46,* 311-349. https://doi.org/10.5735/086.046.0501

Iso-Ahola, S. E. (2017). Reproducibility in psychological science: When do psychological phenomena exist? *Frontiers in Psychology, 8,* 879. https://doi.org/10.3389/fpsyg.2017.00879

Iverson, T. (2014). Generalized fiducial inference. *Wiley Interdisciplinary Reviews: Computational Statistics, 6,* 132-143. https://doi.org/10.1002/wics.1291

Johnstone, D. J. (1987). Tests of significance following RA Fisher. *The British Journal for the Philosophy of Science, 38,* 481-499.

Johnstone, D. J. (1989). On the necessity for random sampling. *The British Journal for the Philosophy of Science, 40,* 443-457.

Klein, S. B. (2014). What can recent replication failures tell us about the theoretical commitments of psychology? *Theory & Psychology, 24,* 326-338. https://doi.org/10.1177/0959354314529616

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams, R. B., Jr., Alper, S.,…Nosek, B. A. (2018). Many Labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science, 1,* 443-490. https://doi.org/10.1177/2515245918810225

Krause, M. S. (2019). Randomness is problematic for social science research purposes. *Quality & Quantity,* 53, 1495-1504. https://doi.org/10.1007/s11135-018-0824-4

Lakatos, I. (1976). Falsification and the methodology of scientific research programmes. In S. G. Harding (Ed), *Can Theories be Refuted?* (pp. 205-259). Springer: Netherlands.

Lakens, D., Adolfi, F. G., Albers, C. J., Anvari, F., Apps, M. A., Argamon, S. E.,...& Buchanan, E. M. (2018). Justify your alpha. *Nature Human Behaviour, 2,* 168–171. https://doi.org/10.1038/s41562-018-0311-x

LeBel, E. P., Berger, D., Campbell, L., & Loving, T. J. (2017). Falsifiability is not optional. *Journal of Personality and Social Psychology, 113,* 254–261. https://doi.org/10.1037/pspi0000106

Lehmann, E. L. (1993). The Fisher, Neyman-Pearson theories of testing hypotheses: One theory or two? *Journal of the American statistical Association, 88,* 1242-1249.

Lehmann, E. L. (1997). Error and the growth of experimental knowledge [Book review]. *Journal of the American Statistical Association, 92,* 789. https://doi.org/10.1080/01621459.1997.10474032

Lindsay, R. M., & Ehrenberg, A. S. (1993). The design of replicated studies. *The American Statistician, 47*, 217-228.

Ludbrook, J., & Dudley, H. (1998). Why permutation tests are superior to *t* and *F* tests in biomedical research. *The American Statistician, 52,* 127-132. https://doi.org/10.1080/00031305.1998.10480551

Macdonald, R. R. (1997). On statistical testing in psychology. *British Journal of Psychology, 88,* 333-347. https://doi.org/10.1111/j.2044-8295.1997.tb02638.x

Mayo, D. (2014). Are *p* values error probabilities? Or, "it's the methods, stupid!" (2nd install). *Error Statistics Philosophy.* Retrieved from https://errorstatistics.com/2014/08/17/are-p-values-error-probabilities-installment-1/

Mayo, D. G., & Cox, D. R. (2006). Frequentist statistics as a theory of inductive inference. *IMS Lecture Notes–Monograph Series: Optimality, 49*, 77-97. https://doi.org/10.1214/074921706000000400

Mayo, D. G., & Spanos, A. (2006). Severe testing as a basic concept in a Neyman–Pearson philosophy of induction. *The British Journal for the Philosophy of Science, 57,* 323-357.

Mayo, D. G., & Spanos, A. (2011). Error statistics. In D. M. Gabbay, P. Thagard, & J. Woods, P. S. Bandyopadhyay, and M. R. Forster (Eds.), *Handbook of philosophy of science: Philosophy of statistics* (Vol. 7, pp, 152-198). New York: Elsevier.

McGuire, W. J. (1983). A contextualist theory of knowledge: Its implications for innovation and reform in psychological research. *Advances in Experimental Social Psychology, 16,* 1-47. https://doi.org/10.1016/s0065-2601(08)60393-7

Meehl, P. E. (1967). Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science, 34,* 103-115. https://doi.org/10.1086/288135

Neyman, J. (1937). X. Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences, 236,* 333-380. https://doi.org/10.1098/rsta.1937.0005

Neyman, J. (1950). *First course in probability and statistics.* New York: Henry Holt.

Neyman, J. (1971). Foundations of behavioristic statistics (with comments). In V. P. Godambe & D. A. Sprott (Eds.), *Foundations of statistical inference* (pp. 1–19). Toronto: Holt, Rinehart & Winston.

Neyman, J. (1977). Frequentist probability and frequentist statistics. *Synthese, 36, 97*-131.

Neyman, J., & Pearson, E. S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference: Part I. *Biometrika 20A,* 175–240. https://doi.org/10.2307/2331945

Neyman, J., & Pearson, E. S. (1933). IX. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society A, 231,* 289-337. https://doi.org/10.1098/rsta.1933.0009

Neyman, J., & Scott, E. L. (1958). Statistical approach to problems of cosmology. *Journal of the Royal Statistical Society. Series B (Methodological), 20,* 1-43.

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences, 115,* 2600-2606. https://doi.org/10.1073/pnas.1708274114

Nosek, B. A., & Errington, T. M. (2017). Reproducibility in cancer biology: Making sense of replications. *Elife, 6,* e23383. https://doi.org/10.7554/eLife.23383.001

Nosek, B. A., & Errington, T. M. (2019). *What is replication?* Retrieved from https://doi.org/10.31222/osf.io/u4g6t

Nosek, B. A., & Lakens, D. (2014). Registered reports: A method to increase the credibility of published results. *Social Psychology, 45,* 137-141. https://doi.org/10.1027/1864-9335/a000192

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science*, *7*, 615-631. https://doi.org/10.1177/1745691612459058

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*, aac4716. https://doi.org/10.1126/science.aac4716

Papineau, D. (1994). The virtues of randomization. *The British Journal for the Philosophy of Science, 45,* 437-450. https://doi.org/10.1093/bjps/45.2.437

Pearson, E. S. (1937). Some aspects of the problem of randomization. *Biometrika, 29,* 53-64.

Pedersen, J. G. (1978). Fiducial inference. *International Statistical Review/Revue Internationale de Statistique*, *46,* 147-170.

Perezgonzalez, J. D. (2015a). Fisher, Neyman-Pearson or NHST? A tutorial for teaching data testing. *Frontiers in Psychology, 6,* 223. doi: https://doi.org/10.3389/fpsyg.2015.00223

Perezgonzalez, J. D. (2015b). The meaning of significance in data testing. *Frontiers in Psychology, 6,* 1293. https://doi.org/10.3389/fpsyg.2015.01293

Perezgonzalez, J. D. (2017). Statistical sensitiveness for the behavioral sciences. *PsyArXiv*. https://doi.org/10.17605/osf.io/qd3gu Retrieved from https://psyarxiv.com/qd3gu/

Perlman, M. D., & Wu, L. (1999). The emperor's new tests. *Statistical Science, 14,* 355-369. https://doi.org/10.1214/ss/1009212517

Rønneberg, L. T. S. (2017). *Fiducial and objective Bayesian inference: History, theory, and comparisons.* University of Oslo. Retrieved from http://urn.nb.no/URN:NBN:no-62990

Rosenberg, M. (1965). *Society and the adolescent self-image.* Princeton, NJ: Princeton University Press.

Royall, R. M. (1997). *Statistical evidence: A likelihood paradigm.* London: Chapman & Hall.

Redish, D. A., Kummerfeld, E., Morris, R. L., & Love, A. C. (2018). Reproducibility failures are essential to scientific inquiry. *Proceedings of the National Academy of Sciences, 115*, 5042-5046. https://doi.org/10.1073/pnas.1806370115

Rubin, M. (2017a). Do *p* values lose their meaning in exploratory analyses? It depends how you define the familywise error rate. *Review of General Psychology, 21,* 269-275. https://doi.org/10.1037/gpr0000123

Rubin, M. (2017b). When does HARKing hurt? Identifying when different types of undisclosed post hoc hypothesizing harm scientific progress. *Review of General Psychology, 21,* 308-320. https://doi.org/10.1037/gpr0000128

Schmidt, S. (2009). Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Review of General Psychology, 13,* 90-100. https://doi.org/10.1037/a0015108

Schneider, J. W. (2015). Null hypothesis significance tests. A mix-up of two different theories: The basis for widespread confusion and numerous misinterpretations. *Scientometrics, 102,* 411-432. https://doi.org/10.1007/s11192-014-1251-5

Schwarz, N., & Strack, F. (2014). Does merely going through the same moves make for a "direct" replication? Concepts, contexts, and operationalizations. *Social Psychology, 45,* 299–311.

Seidenfeld, T. (1979). *Philosophical problems of statistical inference: Learning from R. A Fisher*. London: Reidel.

Senn, S. (2005). Baseline balance and valid statistical analyses: common misunderstandings. *Applied Clinical Trials, 14,* 24-27.

Serlin, R. C. (1987). Hypothesis testing, theory building, and the philosophy of science. *Journal of Counseling Psychology, 34,* 365-371. https://doi.org/10.1037/0022-0167.34.4.365

Shaver, J. P. (1993). What statistical significance testing is, and what it is not. *The Journal of Experimental Education, 61,* 293-316. https://doi.org/10.1080/00220973.1993.10806592

Shrout, P. E., & Rodgers, J. L. (2018). Psychology, science, and knowledge construction: Broadening perspectives from the replication crisis. *Annual Review of Psychology, 69,* 487-510. https://doi.org/10.1146/annurev-psych-122216-011845

Simons, D. J. (2014). The value of direct replication. *Perspectives on Psychological Science, 9,* 76-80. https://doi.org/10.1177/1745691613514755

Simons, D. J., Shoda, Y., & Lindsay, D. S. (2017). Constraints on generality (COG): A proposed addition to all empirical papers. *Perspectives on Psychological Science, 12,* 1123-1128. https://doi.org/10.1177/1745691617708630

Stanley, T. D., Carter, E. C., & Doucouliagos, H. (2018). What meta-analyses reveal about the replicability of psychological research. *Psychological Bulletin, 144,* 1325-1346. https://doi.org/10.1037/bul0000169

Sterba, S. K. (2009). Alternative model-based and design-based frameworks for inference from samples to populations: From polarization to integration. *Multivariate Behavioral Research, 44,* 711-740. https://doi.org/10.1080/00273170903333574

Strack, F. (2017). From data to truth in psychological science. A personal perspective. *Frontiers in Psychology, 8,* 702. https://doi.org/10.3389/fpsyg.2017.00702

Stroebe, W., & Strack, F. (2014). The alleged crisis and the illusion of exact replication. *Perspectives on Psychological Science, 9,* 59-71. https://doi.org/10.1177/1745691613514450

Szucs, D., & Ioannidis, J. (2017). When null hypothesis significance testing is unsuitable for research: A reassessment. *Frontiers in Human Neuroscience, 11,* 390. https://doi.org/10.3389/fnhum.2017.00390

Tajfel, H. (1981). *Human groups and social categories: Studies in social psychology.* Cambridge: Cambridge University Press.

Van Bavel, J. J., Mende-Siedlecki, P., Brady, W. J., & Reinero, D. A. (2016). Contextual sensitivity in scientific reproducibility. *Proceedings of the National Academy of Sciences, 113,* 6454-6459. https://doi.org/10.1073/pnas.1521897113

Venn, J. (1876). *The logic of chance* (2nd ed.). Macmillan and Co.

Veronese, P., & Melilli, E. (2015). Fiducial and confidence distributions for real exponential families. *Scandinavian Journal of Statistics, 42,* 471-484. https://doi.org/10.1111/sjos.12117

Wagenmakers, E. J. (2007). A practical solution to the pervasive problems of *p* values. *Psychonomic Bulletin & Review, 14,* 779-804. https://doi.org/10.3758/BF03194105

Wagenmakers, E. J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J.,...& Matzke, D. (2018). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review, 25,* 35-57. https://doi.org/10.3758/s13423-017-1343-3

Wagenmakers, E. J., & Gronau, Q. (2018, 26th July). *Error rate schmerror rate.* Bayesian Spectacles. Retrieved from https://www.bayesianspectacles.org/error-rate-schmerror-rate/

Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a world beyond "p< 0.05". *The American Statistician, 73,* 1-19. https://doi.org/10.1080/00031305.2019.1583913

Zabell, S. L. (1992). RA Fisher and fiducial argument. *Statistical Science, 7,* 369-387. https://doi.org/10.1214/ss/1177011233

Zuckerman, M., Li, C., & Hall, J. A. (2016). When men and women differ in self-esteem and when they don't: A meta-analysis. *Journal of Research in Personality, 64,* 34-51. https://doi.org/10.1016/j.jrp.2016.07.007

Zwaan, R. A., Etz, A., Lucas, R. E., & Donnellan, M. B. (2018). Making replication mainstream. *Behavioral and Brain Sciences, e120*. https://doi.org/10.1017/S0140525X17001972

Endnotes

1. The Bayesian approach is more similar to the Fisherian approach than it is to the Neyman-Pearson approach. In particular, both the Bayesian and Fisherian approaches are conditioned on the data in hand rather than on a long run of samples. The critical difference between the Bayesian and Fisherian approaches is that the Fisherian approach does not proceed on the basis of the prior probability distribution of the hypothesis under investigation. Indeed, the Fisherian approach requires a "subjective ignorance" about this prior probability distribution because it would allow the recognition of relevant subsets within the population (Fisher, 1959, p. 25-26). Fisher (1959, p. 17, 20-23) argued that, although the Bayesian approach is valid when prior probabilities are known, his significance testing approach is more appropriate when prior probabilities are unknown and, by definition, study-specific prior probabilities are unknown in novel research situations (including non-exact replications).
2. In this article, I focus on the Neyman-Pearson theory (1928, 1933). I do not consider extensions of this theory. For example, I do not consider Berger's (2003) extension of the Neyman-Pearson approach, which considers conditional frequentist testing as a means of unifying the Neyman-Pearson, Fisherian, and Bayesian approaches. In addition, I do not focus on Mayo and Spanos' (2006, 2011) extension, which argues that Neyman-Pearson long run "error probabilities may be used to make inferences about the process giving rise to data, by enabling the assessment of how well probed or how severely tested claims are" (Mayo & Spanos, 2006, p. 328). Importantly, this *error statistical* approach also includes the Fisherian approach. Hence, it represents a "'hybrid' of sorts" (Mayo & Spanos, 2006, p. 333-334). However, it is

unclear how this hybrid is supposed to operate given that the Neyman-Pearson and Fisherian approaches use fundamentally different reference classes for their probability statements (viz., a well-defined objective sample space vs. a hypothetical reference set conditioned on the sample in hand, respectively; Fisher, 1959, p. 78; Lehmann, 1993, p. 1247). It is also notable that the error statistical approach emphasises that the "severity evaluation must be sensitive to the particular outcome" (Mayo & Spanos, 2006, p. 330). This emphasis on the data in hand brings the error statistical approach closer to the Fisherian approach than to the Neyman-Pearson approach (Hurlbert & Lombardi, 2009, p. 326; Lehmann, 1997, p. 789).

3. Note that random sampling is not the same as randomization (Berk et al., 1995; Fienberg & Tanur, 1996; Ludbrook & Dudley, 1998; Papineau, 1994, p. 442). Random sampling refers to the random selection of a sample from a population, and it allows generalization from the sample to the population. Randomization refers to the random assignment of participants to conditions, and it allows a clearer interpretation of research results in the presence of potentially confounding variables. Both Fisher and Neyman and Pearson stressed the importance of randomization in research design (e.g., Fisher, 1937; Neyman, 1950; Pearson, 1937). However, only the Neyman-Pearson approach requires random sampling in order to allow generalization to the objective population (Ludbrook & Dudley, 1998). The Fisherian approach does not require random sampling because generalization is made to a hypothetical sample-specific infinite null population that contains no recognizable relevant subsets (Johnstone, 1989).
4. Some commentators have argued that researchers should abandon significance thresholds (alpha levels) and instead consider *p* values without reference to any benchmark for "significance" (e.g., Amrhein et al., 2019; Amrhein, Korner-Nievergelt, & Roth, 2017; Hurlbert & Lombardi, 2009, p. 318; Wasserstein et al., 2019). This view has sometimes been attributed to Fisher's later publications (Gigerenzer, 1993, p. 316-317; Gigerenzer et al., 2004, p. 11). Although Fisher advised researchers to report exact *p* values in his later publications, he did not advise them to abandon the use of significance thresholds. Instead, he recommended that researchers should not set significance thresholds at "$p \leq .050$" in an automatic fashion, and that they should vary their thresholds in a way that takes into account the particular circumstances of each hypothesis that they test (e.g., Fisher, 1955, p. 74; Fisher, 1959, p. 42, pp. 100-101). In considering the call to abandon significance thresholds, it is important to appreciate that, in the absence of a reference point for determining which *p* values are "significant," "surprising," "small," or "low," *p* values cannot affect researchers' attitudes or guide their behaviour. For example, in the absence of a significance threshold, a *p* value of .0001 does not warrant any action or change in attitude on the part of the researcher because it is quite possible to obtain this *p* value when the statistical null hypothesis is true (Mayo & Cox, 2006, p. 80; Perezgonzalez, 2015b, p. 3). In order for *p* values to contribute to decisions and attitude change, researchers need to imbue them with evidential meaning, and the only way to do this is to interpret them in relation to a significance threshold. The use of significance thresholds enables researchers to make provisional decisions about rejecting the substantive null hypothesis. These decisions then feed into further decisions about conducting follow-up studies and testing new research questions.
5. In the Neyman-Pearson approach, researchers who reject a statistical null hypothesis are not expected to adopt any belief about that hypothesis. They are only expected to act in a way that is consistent with the rejection so that, "in the long run of experience, we shall not too often be wrong" (Neyman & Pearson, 1933, p. 291). In contrast, in the Fisherian approach, researchers

consider pieces of evidence against a substantive null hypothesis that has the potential to "be disproved by a single failure" (Fisher, 1937, p. 19; e.g., "all swans are white"; for a different view, see Hager, 2013, p. 254). This evidence informs the researchers' belief about the substantive null hypothesis via the modus tollens argument. However, it is important to note that a false substantive null hypothesis provides only one potential reason for a significant result. Other reasons include the various statistical and methodological assumptions that form part of the overall null model, including distributional assumptions and systematic errors (Greenland & Chow, 2019). Hence, a significant result may be due either to a false null hypothesis or to a false assumption in the null model or both. In interpreting significant results, researchers need to weigh up the likelihood of each of these potential explanations in the context of a priori theory and evidence, robustness checks, error checking, and logical reasoning in order to arrive at a provisional decision about whether the result is better explained by a false null hypothesis or by a false assumption. For example, they must weigh up the likelihood that a significant result was primarily caused by a problematic violation of a distributional assumption, a systematic data coding error, or a false null hypothesis (Fisher, 1959, pp. 39-41). This process of inference to the best explanation may be informed by the results of tests of distributional assumptions and checks for coding errors as well as a priori knowledge about the likelihood of such assumption violations. Researchers can then make an informed decision about whether or not to conclude that the significant result was caused by a false null hypothesis rather than a false statistical or methodological assumption. Given that researchers can err during this process of inference to the best explanation, as well as the potential for Type I errors, their decisions are only "provisional" (preliminary), pending further corroboration.

6. Taking into account the fact that significant results may arise from "chance coincidence" (i.e., Type I errors), Fisher (1937, p. 16) argued that "no isolated experiment, however significant in itself, can suffice for the experimental demonstration of any natural phenomenon." Researchers are only able to make a firm decision about rejecting a substantive null hypothesis on the basis of a series of real direct and conceptual replications that "rarely fail to give us a statistically significant result" (Fisher, 1937, p. 16; see also Fisher, 1926, p. 85, p. 504). Hence, a single *p* value from a single study provides only a "provisional" piece of evidence against the substantive null hypothesis. The word "provisional" indicates that "no irreversible decision has been taken; that, as rational beings, we are prepared to be convinced by future evidence that appearances were deceptive, and that in fact a very remarkable and exceptional coincidence had taken place" (Fisher, 1959, p. 35; see also Fisher, 1959, pp. 100-101).
7. Some commentators have argued that Fisherian *p* values cannot be equated with Type I errors (e.g., Hubbard, 2004, 2011; Hubbard & Bayarri, 2003). In making this argument, they compare the Fisherian *p* value with the Neyman-Pearson alpha level. I agree with this line of reasoning. In the present article, I make a separate argument by comparing the Fisherian significance threshold (e.g., $p \leq .050$) with the Neyman-Pearson alpha level (e.g., $\alpha \leq .050$), and I am careful to distinguish the different implications of passing this threshold in each case. Hence, consistent with Hubbard and colleagues, I do not equate a Fisherian *p* values with the Neyman-Pearson alpha level.
8. Fisher's fiducial argument provides a method of making a probability statement about a population parameter conditional on a sample statistic in the absence of prior information about the parameter. This *fiducial inference* has been the subject of much controversy and criticism, including the concern that the resulting fiducial probabilities are not additive (for reviews, see

Fraser, 2008; Zabell, 1992). At one point, fiducial inference was described as being "essentially dead" (Pedersen, 1978, p. 147). Contrary to this view, there has been a renewed interest in fiducial inference, with several statisticians currently investigating various instantiations of this idea (e.g., Bowater, 2017; Hampel, 2003; Heike et al., 2007; Iverson, 2014; Veronese & Melilli, 2015; for reviews, see Hannig, Iyer, Lai, & Lee, 2016; Rønneberg, 2017). It is beyond the scope of the current paper (and the competency of its author) to consider the fiducial argument in detail. It is sufficient to note that Fisher's epistemic view of probability allows the consideration of a sample-specific significance testing approach (i.e., fiducial probability statements about sample statistics that are conditioned on hypothetical infinite null populations that contain no recognizable relevant subsets) without the need to consider probability statements about population parameters that are conditioned on hypothetical superpopulations that contain no recognizable relevant subpopulations (i.e., fiducial inference; for a similar conclusion, see Lehmann, 1993, p. 1242).

9. The Neyman-Pearson Type II error rate refers to the maximum frequency of incorrectly rejecting the statistical alternative hypothesis in the case of a long run of exact replications that randomly draw different samples from the same objective population. This error rate refers to a single precise statistical alternative hypothesis, its associated "true" effect size in comparison with the null hypothesis, and the test's power to detect that effect (Perezgonzalez, 2015a, 2017). The Fisherian approach does not endorse any of these concepts. Nonetheless, Fisherian researchers do consider multiple possible alternative substantive hypotheses (models, populations) in contrast to the substantive null hypothesis (e.g., Fisher, 1959, p. 35, pp. 78-79), and Fisher discussed the *sensitivity* of tests instead of their power (e.g., Fisher, 1937, pp. 25-26; Gigerenzer, 1993, p. 320; Hubbard & Bayarri, 2003, p. 173; Hurlbert & Lombardi, 2009, p. 318; Lehmann, 2011, p. 51; Lehmann, 1993, p. 1245; Macdonald, 1997, p. 339; Meehl, 1967, p. 107; Perezgonzalez, 2015a, 2017). Furthermore, Perezgonzalez (2017) has recently shown that it is possible to compute a priori and post hoc sensitiveness for Fisherian tests if researchers have an idea of the minimum effect size in which they are interested. In this approach, sensitiveness is defined as the minimum sample size that is required in order to obtain a significant result at a specified significance threshold for a given minimum effect size of interest.

Acknowledgements

I am grateful to the following people for their comments and criticisms on earlier drafts of this article: Sander Greenland, Brian Haig, Julian Marewski, Jose Perezgonzalez, Stephen Senn, and David Trafimow. I am also grateful to Giedrius Trakimas for discussions that caused me to clarify Endnote 4.

Funding

The author declares no funding sources.

Conflict of Interest

The author declares no conflict of interest.