Journal of the
Experimental Analysis of Behavior

JOURNAL OF THE EXPERIMENTAL ANALYSIS OF BEHAVIOR 2019, **111,** 329–341 NUMBER 2 (MARCH)

Replicability and randomization test logic in behavior analysis

KENNETH W. JACOBS

UNIVERSITY OF NEVADA, RENO

Randomization tests are a class of nonparametric statistics that determine the significance of treatment effects. Unlike parametric statistics, randomization tests do not assume a random sample, or make any of the distributional assumptions that often preclude statistical inferences about single-case data. A feature that randomization tests share with parametric statistics, however, is the derivation of a *p*-value. *P*-values are notoriously misinterpreted and are partly responsible for the putative "replication crisis." Behavior analysts might question the utility of adding such a controversial index of statistical significance to their methods, so it is the aim of this paper to describe the randomization test logic and its potentially beneficial consequences. In doing so, this paper will: (1) address the replication crisis as a behavior analyst views it, (2) differentiate the problematic *p*-values of parametric statistics from the, arguably, more useful *p*-values of randomization tests, and (3) review the logic of randomization tests and their unique fit within the behavior analytic tradition of studying behavioral processes that cut across species.

Key words: counterfactual reasoning, general process approach, null hypothesis significance testing, randomization tests, replication crisis, single-case experimental designs, statistical inference

The purpose of the present essay is to describe, as plainly as possible, the logic of a statistical test that determines the significance of treatment effects. That statistical test is called a randomization test (RT), and unlike the statistical tests of significance that are parametric, RTs are readily applicable to single-case experimental designs (Edgington & Onghena, 2007). Because the application of RTs to single-case data has been elucidated (Edgington, 1980; 1987) and tutorials have been provided (Heyvaert & Onghena, 2014; Onghena & Edgington, 2005), the current paper is only concerned with the logic of RTs and the consequences that follow from it.

"Plainly as possible," the qualifier in this essay's purpose, is meant to engage the part of behavior analysis that is practical. The aim is to afford effective action on the part of readers, especially if those readers are fellow behavior analysts with no formal training in statistics. Skinner (1956) and his contemporaries (e.g., Sidman, 1960) were right to reject certain statistical methods (e.g., group averages), but it seems they may have "thrown the baby out with the bathwater." Their haste may even be responsible for the generation of behavior analysts that reject statistical methods out of hand. I must admit that behavior analysis was appealing because "you didn't need any of that stats stuff." As cited in Moore (1995), "…you can learn Skinnerian psychology very quickly: The first day you are there, you learn that statistics is no damn good. Bang! like that" (Baars, 1985, p. 206). Like that behavior analysts get siloed.

One class of statistical tests applicable to single-case data is the nonparametric RT pioneered by E.J.G. Pitman. Whereas parametric statistics make assumptions about population parameters to determine the significance of data sampled, Pitman's (1937a; 1937b; 1938) nonparametric statistics determine the significance of data based on how the experimental units in a study are organized. For instance, there are 252 different ways to *randomize,* or randomly assign, 10 participants equally among two treatments: 10!/5!5! = 252. When only one of these ways makes up the actual experiment conducted, there is an opportunity for the researcher to ask: "What if I had selected a different way to randomize the participants among treatments? If participants from one treatment group were exposed to the other treatment and vice versa, would I have obtained the same results?" RTs answer such questions by comparing the test statistic

Kenneth W. Jacobs, Department of Psychology, University of Nevada, Reno.Thank you to Peter Killeen and Martha Zimmermann for their helpful comments on earlier drafts.

Correspondence concerning this article should be addressed to Kenneth W. Jacobs, Department of Psychology/296, University of Nevada, Reno, NV 89557. Email: kjacobs789@gmail.com

doi: 10.1002/jeab.501

value (e.g., mean difference) of participants' observed scores against the test statistic values that could have arisen if participants were randomized according to the other, 251 possible assignments to treatments. RTs, therefore, allow one to ask if there is a significant difference between treatment scores based on the way participants were randomized among those treatments.

RTs are unique among the other statistical methods available because they come with very few assumptions and their inferences do not extend beyond that which is immediately observed. RTs do not require any reference to a population besides the one sampled, so the inference is about the participants in the study and their behavior under various conditions in that study. This type of inference is in contrast to the parametric statistical inference about a population of individuals neither in the experiment conducted nor exposed to the treatments administered. Statistical inference about the very participants studied, rather than about populations unknown, is one of the features that should make RTs attractive to behavior analysts. The feature of RTs that might make behavior analysts wary, however, is the feature they share with parametric statistics: a probability value known as the *p*-value.

P-values are an index of statistical significance and are the heart of what is called null hypothesis significance testing (NHST). Simply put, NHST is when a researcher attempts to reject the null statement that there is no difference between treatments in the assumed population. The index for rejecting (or failing to reject) a null is the *p*-value (e.g., $p < .05$), the probability of the results of an experiment or more extreme given the null hypothesis. RTs may be considered a form of NHST because they return a *p*-value and require a null hypothesis. An unfortunate consequence of this formal similarity is that the *p*-values of traditional, parametric NHSTs are the center of controversy concerning the replicability of psychological science research (Branch, 2018). NHST *p*-values are notoriously misinterpreted and have been named culprit of the putative "replication crisis" (Killeen, 2018; see Open Science Collaboration, 2015; see Pashler & Wagenmakers, 2012, for a brief overview of the crisis). Because RTs may be considered a type of NHST, it is prudent that this essay takes the "crisis" into consideration. It would be unfortunate if the RTs useful to behavior analysts got lumped into the tradition of NHST that is predicated on impractical, parametric assumptions.

Nonreplicability is not a Crisis

"Crisis" does not appropriately characterize the nonreplicability of psychological science research. If we take the first definition of crisis in the *New Oxford American Dictionary* (Stevenson & Lindberg, 2015), "a time of intense difficulty, trouble, or danger", then we might agree that replication is hard with few rewards, is annoying when it does not work, and is a threat to the theories that organize how we speak about our observations. These may be some of the personal difficulties scientists face when they fail to replicate results, but science—psychological or otherwise—is not in intense trouble or danger because of replicability per se. Science has its difficulties because of the fallible claims made by the organisms producing it. As Hayes (1991) put it:

We are not testing the consistency of the universe when we replicate research. If we did exactly the same thing in every detail, the same results would occur. Rather, our purpose is to see whether doing what the author said is doing the same thing. We are testing the functional adequacy of the researcher's verbalizations in guiding our behavior (p. 419).

What an author says, however, is under the control of more than that which any "pure" scientific investigation entails. There is that which the author did and observed in the laboratory, but there is also that which the author brings to the laboratory in the form of cultural presuppositions (Kantor, 1953). The prospect of grant funding for tenure and promotion could very well be a source of control that distorts what authors say (Lilienfeld, 2017), which in turn, renders the verbalizations of those researchers functionally inadequate in guiding our behavior as scientists and even citizens. No claims made are infallible, so what is called a crisis today has always—as long as humans have been reporting on their interactions with the environment—been a cause for concern, contemplation, and further inquiry.

The second definition of crisis gets us closer to the current state of replicability in psychological science, for this is "a time when a difficult or important decision must be made."

Indeed, this is a time for important decisions to be made about the methods we value and the creeping assumptions many NHST researchers left unattended (see Gigerenzer, 2004). Decisions, however, are not born out of crises limited to replication alone. Difficult decisions are made in science daily: They range from methodological to societal, especially in behavior analysis (Baer, Wolf, & Risley, 1968; Mace & Critchfield, 2010). Furthermore, the decisions we make as individuals within a scientific community are *indefinitely prolonged* (R.A. Putnam, 2002, referencing C. S. Peirce). Science is a continuous mode of inquiry that does not end with nonreplicability, nor even with a single replication. There will be continuing need for research on important topics, which brings us to the third and final definition of crisis.

The third listed definition of crisis, "the turning point of a disease when an important change takes place, indicating either recovery or death", portrays nonreplicability as analogous to an illness. It incorrectly dichotomizes the problem of replicability as a matter of life or death for the psychological sciences. The prospect that psychological science could die makes at least two presumptions: (1) that its proponents would disband and (2) that "psychology" is not the answer to our human problems. The two are inextricably linked and the putative death of behaviorism by the "cognitive revolution" is an example: (1) behavior analysts did not disband because (2) behavioral research was and continues to be warranted by human problems such as Autism Spectrum Disorder, drug and alcohol abuse, overeating, and gambling. Psychological science—cognitive, behavioral, or both—has much to offer and to investigate in the way of reducing human suffering (Hofmann & Hayes, 2018). Nonreplicability, then, does not mean death to the science driven by our human interests and warranted by our human needs. Nonreplicability entails further inquiry into the methods of the science (e.g., NHST) that have misguided the scientists within that community. It was by way of inquiry that people learned from "the failure of such methods as the method of tenacity, the method of authority, and the method of appeal to allegedly a priori reason" (H. Putnam, 2002, p. 22). Today, we too can learn from the failure of the method of NHST to effectively guide our research endeavors.[1]

NHST is Problematic ($p < .05$)

When behavior analysts learn of the putative replication crisis, it comes as no surprise that "defective contingencies of reinforcement" (Skinner, 1971, p. 470) are central to the issue of replicability. Even nonbehavioral scientists were apt to recognize this when they said: we need to "adjust for human cognition," hold "researchers accountable for analyses," and the "norms of practice must be changed from within" (Leek et al., 2017, pp. 557-559). Their qualms pertain largely to the misinterpretation of *p*-values associated with NHST, which Jacob Cohen (1995) addressed when he said: "Incidentally, I do not question the validity of NHST, but rather its widespread misinterpretation" (p. 1103). Today the argument is close to the same, as demonstrated by Saltelli and Stark's (2018) statement that "…there is nothing technically wrong with *P* values. But even when they are correct and appropriate, they can be misunderstood, misrepresented and misused—often in the haste to serve publication and career" (p. 281). Misuse in service of "publication and career" adds to the NHST conversation by putting more emphasis on the problem that is human behavior under the control of contingencies other than what was immediately observed in any investigative situation.

That RTs return a *p*-value poses this challenge: Can we have a *p*-value without the convoluted logic one must maneuver in order to correctly interpret it (Cohen, 1994); and can we have a *p*-value without the misuses due in part by career-oriented claims? The answer to this question, arguably, is yes. There is nothing technically wrong with *p*-values per se. The problem is with the assumptions that underlie traditional NHST. The heterodoxy of RTs makes the accurate interpretation of *p*-values, and conservative claims based on them,

[1]R.A. Fisher appeared to be aware of this potential failure when, in the *Proceedings from the Society for Psychical Research,* he stated: "The test of significance only tells him what to ignore, namely all experiments in which significant results are not obtained. He should only claim that a phenomenon is experimentally demonstrable when he knows how to design an experiment so that it will rarely fail to give a significant result" (1929, p. 191).

possible and likely because RTs are not based on the most difficult and seldom satisfied assumption in statistics: a random sample of the population to which generalization is to be made.

Differentiating RTs from Traditional NHST

In their guide on the misinterpretations of *p*-values and other statistical tools, Greenland et al. (2016) emphasized what is left out of most traditional definitions of the *p*-value: "In logical terms, the P value tests *all* the assumptions about how the data were generated (the entire model), not just the targeted hypothesis it is supposed to test (such as a null hypothesis)" (p. 339; emphasis theirs). This is to say that the *p*-value is only interpretable with confidence in those rare cases when no assumptions have been violated. The first and foremost assumption rarely met in NHST is the assumption that participants are randomly drawn from a population to which generalization will be made. This is the assumption of a random sample, which supposes that the participants in a study had an equally probable chance of being selected from the population of interest, and therefore, are representative of that population. Since most traditional, parametric NHSTs rely on a random sample to make inferences, the "entire model" Greenland et al. refer to can be called the random sample model (see Onghena, 2018a). Any interpretation of a *p*-value based on the random sample model depends on whether the experimenter randomly sampled participants from a population.

The untenable random sample. The most basic and conspicuous assumption in NHST—that participants were sampled at random from a population of interest—is untenable because it is so rarely achieved. It is an assumption often taken for granted, and one need look no further than their university setting to find evidence of this behavior instilled in the notion of a "convenience sample." Introductory psychology students are usually the subjects of those convenience samples, more aptly described as biased samples. Inferences about the larger population of interest, in the case of biased samples, are not warranted. However, authors make those inferences regardless, either explicitly in their discussions or implicitly when they write "$p < .05$". The criteria by which consumers must judge those inferences begin with the question: Did the authors obtain a random sample?

In settings outside the university, applied researchers quickly discovered the untenable nature of the random sample model. Hayes, Barlow, and Nelson-Gray (1999) distinguished between the "conceptual population" and the "known population" (p. 23). The identification of *all* the characteristics relevant to a particular behavioral disorder of interest (e.g., drug abuse defined by a diagnostic manual) constitutes the conceptual population. The identification of *all* the clients that have the characteristics of the conceptual population constitutes the known population. The challenge that Hayes et al. pointed out is that conceptual and known populations are not isomorphic: Various persons that meet the criteria of the conceptual population do not seek treatment, and as a result, are not known. How, then, do we sample persons that do not seek treatments, do not have contact information, do not have a phone or computer, do not have a home, or have too many homes to be on our radar? A fully known population is difficult, if not impossible, to attain and sample (Dugard, 2014).

The last line of defense for those in favor of the random sample model is *representativeness*. The logic is that as long as researchers select enough participants reflective of the larger population, they will be able to make warranted and valid inferences about that larger population (i.e., even though they did not randomly sample from that population). In an encyclopedic entry on inference, Zelen (1998) stated: "In what follows it will *always be assumed* that there is both a well-defined population and a data collection plan which does not create opportunities for systematic error" (p. 2035, emphasis added). The well-defined population is akin to the Hayes et al. (1999) conceptual population while the data collection plan is akin to the random sampling of participants from the known population. Representative samples may be well-defined, but they are not the random samples required for unbiased inferences about the population of interest. Additionally, the parametric t and F tables, from which researchers determine their *p*-values, are predicated on the

assumption that the experimenter selected a random sample. "Parametric statistical tables are applicable only to random samples, and their invalidity of application to nonrandom samples is widely recognized" (Edgington & Onghena, 2007, p. 6). To use them with convenience samples is to abuse them.

The random assignment model. RTs are not based on the random sample model. Instead, they are predicated on the random assignment model, which supposes that an experimenter randomly assigned participants, or other experimental units like measurement times, to treatments. Unlike random sampling, random assignment is practically feasible and largely under the experimenter's control.

Basis in a random assignment model means RTs are free from the assumption of a random sample *and* every other assumption that comes with it: normality of distribution of the residuals, homogeneity of variance, independence of observations, etc. In the words of Edgington and Onghena (2007): "In a sense, randomization tests are the ultimate nonparametric tests. To say that they are free from parametric assumptions is a gross understatement of their freedom from questionable assumptions—they are free from the most conspicuously incorrect assumption of all, which is the assumption that the subjects or other research units were randomly drawn from a population" (p. 1). The random assignment of participants to treatments introduces probability into an experimental design and allows one to make statistical inferences about what could have happened if those assignments to treatments were different. Rather than say something about the larger population, RTs say something about what you did as an experimenter and observed as a result. The random assignment model forgoes the fatal attraction of making universal statements about populations unseen, in favor of making defensible statements based on data in hand.

That RTs do not and cannot make statistical inferences about the larger population is a strength rather than a weakness. Statistical inferences made on the basis of convenience samples, representative samples, and other nonrandom samples are biased and invalid when extended beyond the convenience samples on which they are based. RTs have the advantage of making statistical inferences—about the data immediately in hand—when participants are not randomly sampled (Edgington, 1966). For example, applied researchers can nonrandomly sample from a population of drug abusers and experimentalists can nonrandomly sample from a population of undergraduates. A well-defined conceptual population, in regard to your investigative goals, is still important but does not come with the near-impossible task of randomly sampling from an ill-known population. The only stipulation is that researchers must perform a random assignment for any RT to be valid (Onghena, 2018b).

Traditionally, NHSTs such as analysis of variance (ANOVA) and *t*-tests are based on the random sample model. Alternative test statistics equivalent to the ones used in ANOVA and *t*-tests, though, are readily available when participants are randomly assigned to treatments. Edgington and Onghena (2007) describe those tests, and more, while also providing a means to run them with software for any Windows-based computer (see Huo & Onghena, 2012, https://ppw.kuleuven.be/mesrg/software-and-apps/rt4win). When traditionally parametric NHSTs—like ANOVA and *t*-tests—are based on the random assignment model, they have the advantage of making valid statistical inferences without the hassle of a random sample. Additionally, the versatility of the random assignment model to guide the creation of not just equivalent test statistics, but new ones as well, means that there are RTs available for the equally versatile design elements of single-case experiments (Edgington & Onghena, 2007). Behavior analysts sometimes rely on nonparametric rank tests such as the Wilcoxon-Mann–Whitney test or the Kruskal-Wallis test to determine the significance of their effects within subjects, but these tests suffer a loss of precision when observed scores are transformed into ranks (Onghena & Edgington, 2005). Equivalent RTs, or RTs tailored for use with single-case experimental designs, are readily available and do not degrade one's raw data by transforming it into ranks (see Edgington & Onghena, 2007).

Randomization Test Logic

The distinction between the random sample model and random assignment model is an important one; it differentiates RTs from the traditional NHST that has posed problems for

the basic and applied researchers that rely on them. The random sample and assignment distinction is also important for terminological reasons, as RTs are often confused with permutation tests and various bootstrapping techniques. RTs are neither types of permutation tests nor types of bootstrapping tests because RTs have their basis in the random assignment model while permutation and bootstrapping tests have their basis in the random sample model (see Onghena, 2018a, for a review). These distinctions are important because *random assignment* is the raison d'être of all RT logic.

What's the Logic?

The RT logic is a series of operations that begins with the practical organization of materials—randomization—and ends with a proposition based on statistical inference. As is the case with any statistical test, the RT logic could be misunderstood, misrepresented, or misused. For instance, researchers could defy the RT logic by overgeneralizing the implications of their results. A redeeming quality of the RT logic, however, is that it is straightforward compared to traditional NHST. In Edgington and Onghena's (2007) words: "Unlike conventional parametric tests, randomization tests determine P-values directly from experimental data without reference to tables based on infinite continuous probability distributions, so the procedure is easily understood by persons without a mathematical background" (p. 13). If the RT logic does not require a mathematical background to be understood, then the communities of researchers and practitioners using RTs could be in a better position to guard against misunderstandings, misrepresentations, or misuses. Eschewing misuses such as overgeneralization may begin with an understanding of the series of operations entailed in randomization.

The operations entailed. Randomization, or random assignment in particular, is the necessary requisite for conducting an RT. Randomization is a definite activity with definite materials worked upon. Below is a thumbnail of the operations involved in randomization and the more abstract operations that follow from it. Because the form of randomization familiar to most researchers is the random assignment of participants to treatments in a group design, terminological clarification is required for what follows.

The series of operations listed below apply to a single-case experimental design (SCED), where the experimental units of interest are the responses of a single subject rather than the responses of multiple subjects within different groups. In the case of SCEDs, randomization involves the random assignment of treatments (e.g., treatments A and B) to *measurement times*. 'Measurement times' refer to the frequent and repeated measurements of an individual's behavior for an extended period or periods of time. While SCED elements—withdrawal, multiple baseline, and alternating treatments—inform the frequency with which a scientist or practitioner will measure the behavior of interest, random assignment determines *when* to take those measurements. For simplicity of exposition, the following operations apply to an alternating treatments design (ATD; Barlow & Hayes, 1979) testing the effects of treatments A and B (cf. Heyvaert & Onghena, 2014; Onghena & Edgington, 2005; see Huo & Onghena, 2012, for a review of the group design logic).

1. Decide on the number of measurement times for each treatment based on your research question and practical feasibility. For example, twenty measurement times (sessions or trials) divided equally between two treatments (A and B). Note that the statistical power of any single-case RT is partly determined by the number of measurement times (Onghena & Edgington, 2005). Generally, the more measurement times the better, but there are some exceptions (see Heyvaert & Onghena, 2014).
2. Determine the number of consecutive presentations of the same treatment allowable. Presenting Treatment A 10 times in a row and Treatment B 10 times in a row might increase the likelihood of carry-over effects. The maximum number of consecutive treatments that may minimize carry-over effects, for the sake of this example, is three. That is, treatments A and B can be presented no more than three times consecutively.
3. Given the parameters of your ATD—20 measurement times divided equally

between two treatments with a limit of three consecutive presentations of the same treatment—calculate the number of possible ways to assign treatments A and B among the 20 measurement times. Use the R Studio package developed by Bulté and Onghena (2008) or the Shiny Single Case Data Analysis (Shiny SCDA) web application (https://tamalkd.shinyapps.io/scda/). The number of possible assignments based on the Shiny SCDA is 66,486. There are 66,486 ways to assign treatments among the 20 measurement times.

4. Among the number of possible assignments—66,486—you then need to choose one at random. Using the Shiny SCDA we get the random assignment shown in Table 1. Note that treatments are organized with respect to measurement times and no treatment is presented more than three times consecutively.
5. Choose a test statistic best suited for determining whether there is a difference between treatments. Suppose your dependent measure is percentage correct on an academic task and you are comparing the effectiveness of one treatment (A) versus another treatment (B). The test statistic you choose is the absolute difference between treatment A and B means: |A-B|. Note that RTs allow for the use of many different test statistics.
6. Implement your treatments based on the random assignment chosen. Suppose the data you collected is that shown in Figure 1. Visual analysis shows that there is a difference between treatments A and B, but more evidence is needed to determine if that difference is statistically significant.
7. Calculate the mean percentage correct for each condition and calculate the difference between those means to obtain the test statistic described in Step 5. The absolute difference between treatment means is: 0.30. This can be obtained using Shiny SCDA.
8. Use Shiny SCDA to construct a *reference set* composed of all of the 66,486 possible assignments. Or, for ease of computation, construct a reference set by randomly selecting at least 1,000 of the possible

Table 1

Random assignment of treatments to measurement times and percentages correct on an academic task

	Measurement Times																			
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Treatment	A	B	A	B	B	A	B	A	B	B	B	A	A	B	A	B	A	A	A	B
Response	87%	60%	85%	55%	58%	90%	62%	83%	56%	60%	54%	90%	88%	61%	85%	53%	86%	90%	90%	55%

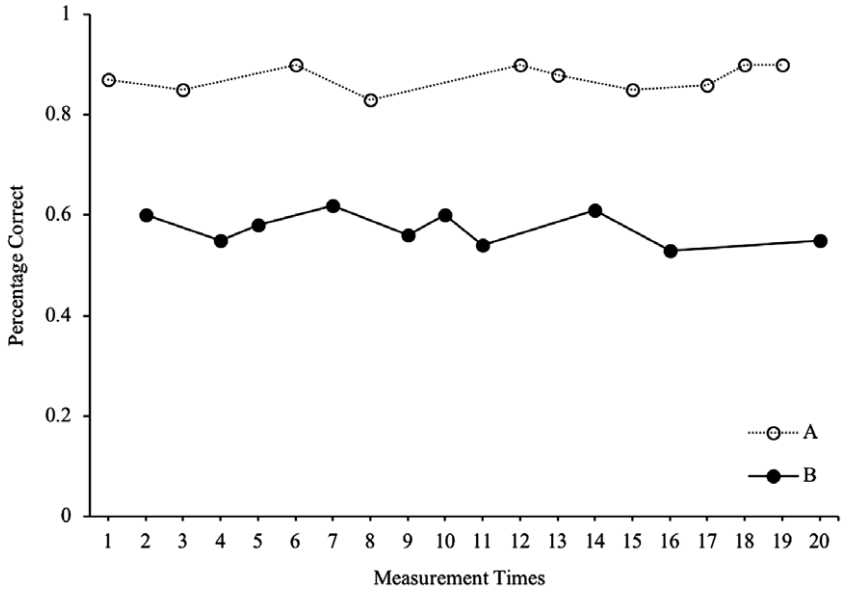

Fig. 1. Hypothetical percentages correct obtained per treatment in a randomized alternating treatments design.

assignments. The former constitutes an exhaustive/systematic RT while the latter constitutes a nonexhaustive Monte Carlo RT. The Monte Carlo RT is more feasible in this case, as it is less computer-intensive and will not take as long to do in Shiny SCDA. Note that as few as 1,000 randomly selected assignments has been cited as the minimum for valid, nonexhaustive tests of statistical significance (see Edgington & Onghena, 2007, for discussion over Monte Carlo RTs).

9. The purpose of the reference set is to superimpose the obtained percentages correct for each treatment over all of the other assignments to determine how many times the same test statistic or larger would be obtained. Shiny SCDA will construct the reference set, superimpose the percentages correct over the other assignments, calculate the new absolute differences between treatment means, and return the number of times the same test statistic or larger was obtained. The number of times the same test statistic or larger was obtained using the Monte Carlo RT with 1,000 assignments was 1.[2]
10. Calculate the *p*-value. If there is only one assignment with a test statistic that is equal to or larger than the test statistic you actually observed, then your *p*-value is: $p = .001$. This is to say that the *p*-value is the proportion of the 1,000 possible assignments that have a test statistic equal to or larger than the observed test statistic. Hence, 1/1,000 gives .001.

The 10 operations outlined above are not exhaustive. They comprise the key activities necessary for running an RT. An important element left out of this series of operations is any mention of the *null hypothesis*. Its exclusion from the 10 steps is because it is less of an operation and more of an assumption that gives those operations meaning. The generic null hypothesis, stated as "no difference between treatments," is the conditional on which we interpret the *p*-value.

The null (H_0). The null hypothesis—no difference between treatments—assumes that any

[2]If you were to run the Monte Carlo RT with a reference set of 1,000 assignments a second time, it might return a different number because the 1,000 assignments randomly selected could be different than the ones selected originally. This is not a cause for concern, as the Monte Carlo RT *p*-value should fall within a 99% probability interval associated with the exhaustive/systematic RT *p*-value (Edgington & Onghena, 2007).

response measured would have been the same regardless of treatments. The null takes it for granted that any of the 66,486 assignments we could have selected would result in the same test statistic or larger. The assumption of the H_0 is the condition on which the *p*-value of any RT is interpreted (see Edgington & Onghena, 2007, for nulls that specify conditions other than just "no difference between treatments"). Put simply, the *p*-value is the probability of the obtained data given the null, which can be expressed in mathematical form as: $p(D|H_0)$.

The *p*-value we obtained using the hypothetical data in Figure 1 was .001, which means there was 1/1,000 possible assignments that resulted in the same test statistic observed or larger. It can be said that the test statistic obtained—in the context of the assignment we randomly chose—is unlikely given the null (no difference between treatments). If the RT had returned $p = 1$, or 1,000/1,000, then the conclusion drawn would be that the test statistic obtained is likely given the null (no difference between treatments). At this juncture, researchers might be inclined to make inferences based on the respective *p*-value obtained, but any inference made involves another series of operations that require explication. The operations involved in making an inference are more abstract but inextricably related to the randomization executed at the outset. The inferential operations that follow from a *p*-value are based on the relations between the assignment chosen at random and the results obtained.

What's the Inference?

When researchers arrive at their $p < .05$ value, they want to say something about the "falsity" of the null hypothesis and "truth" of their alternative hypothesis. Authors want to say something about people in general, or about the size of an effect based on the *p*-value. They want to use the *p*-value to draw conclusions about the probability of their hypotheses, probability that the data will replicate, or probability that their data were due to chance alone. The operations involved in obtaining a *p*-value, however, do *not* warrant any of the above claims (Greenland et al., 2016; Wasserstein & Lazar, 2016). We can make inferences, but they must be based on the operations performed. If our claims are based on the operations involved in obtaining the *p*-value, then our claims have, as John Dewey (1941) would say, *warranted assertibility*.

The inference is statistical. RT inferences are based on the relations between our null assumption, the randomization we engendered, and the data we observed. Our statistical inference, therefore, is based on the following operations.

1. The act of randomization introduces probability into our experimental design, as there was an equally probable chance of selecting any of the 66,486 possible assignments.
2. Assume that any response measured would have been the same regardless of treatments.
3. Operations 1 and 2 allow us to ask a counterfactual: "If we had randomly chosen a different assignment, would the test statistic really be the same regardless of the treatments to which the responses measured are assigned?"
4. We test operation 3 by constructing a reference set and determining the number of test statistics that are equal to or larger than the test statistic observed with our original assignment. If the proportion of test statistics equal to or larger than the one observed is less than .05, we say these data are unlikely given the null (no difference between treatments).
5. When there are very few assignments (e.g., 1/1,000) that result in the same test statistic or larger, we make the inference that one or more of the individual's responses must have been different *because* of the nature of the treatments.

Operation 5 is the statistical inference, and it is warranted and valid based on the operations before it. Note that the inference is not about people in general, is not about the probability of our hypothesis, and is not about the probability that our hypothesis is true. Our statistical inference is based on the randomization we engendered, assumption we made, and data we observed with respect to each treatment. Furthermore, our statistical inference is a causal claim in that it follows from counterfactual reasoning (Onghena, 2018b).

The inference is also counterfactual. Counterfactual reasoning is one of the ways we make causal claims (Holland, 2005). Take the everyday occasion on which you state the counterfactual that is: "If I had said that differently, we wouldn't be having this argument right

now." This counterfactual statement is based on the fact that you said provocative statement *X*, which caused argument *Y*. The reasoning is that if you had not said *X*, you would not have caused *Y*. If the counterfactual is correct in that you do not get *Y* without *X*, then you can more confidently say that *X* causes *Y*.

In the case of our hypothetical results shown in Figure 1 our counterfactual is this: "If we had chosen a different assignment at random, would the test statistic results be the same?" The RT logic assumes that there is no causal relation between the percentages correct we measured and the treatments we administered. On this assumption it would be the case that the responses measured in terms of percentages correct would be the same regardless of the treatments to which they are assigned. By taking our obtained percentages correct and assigning them to different treatments based on the 66,485 other possible assignments, we can test whether we would get the same result. If we do get the same result, say 1,000/1,000 times, then we infer that the nature of the different treatments was not the cause of our responses measured. If we find out that the results are not all the same, say 1/1,000, then we conclude with an inference based on counterfactual reasoning: RT logic assumes that the obtained percentages correct were not caused by the treatments. However, when we assign the obtained percentages correct to the other possible assignments in our reference set, we do not get the same result. Therefore, the nature of the treatments, rather than something else, must be playing a causal role in the obtained percentages correct. If the test statistic result is not the same across all possible assignments, then something about the treatments must have caused one or more of the percentages correct to be different.

By introducing probability into experimental designs through randomization, and by assuming that the responses measured are not due to treatments, we can make inferences that are statistical on account of randomization and causal on account of counterfactual reasoning. We ask, "What if the assignment had been different?" and test that counterfactual by constructing a reference set and superimposing our data onto other possible assignments. If we follow the RT logic and arrive at $p < .05$, then we can make the inference that the difference in treatments caused a difference in one or more of the responses we observed and measured.

The inference does not say it all. An important caveat applies to the causal inference that RTs make: Counterfactuals are an *indirect verification* in that they refer to what could have happened if some condition were absent or different. We test this by superimposing our obtained data onto many other possible assignments in a given reference set, but those were not experiments that we actually conducted. That we did not expose a subject to all possible assignments is what is meant by indirect verification. The statistical and counterfactual claims made are valid and warranted based on the RT operations performed, but to function as guides for everyday human affairs, those claims must be *directly verified* across different assignments, different organisms, and in the laboratory and in practice.

In addition to the inferences made based on RT *p*-values, behavior analysts using single-case experimental designs (SCED) have the advantage of relying on a host of other indices to make their claims: differences in level and trend, percentage of nonoverlapping data (PND), and effect-size measures (Shadish et al., 2014). Furthermore, SCED elements control for many of the confounding variables related to the frequent and repeated measurement of behavior over extended periods of time (Perone & Hursh, 2013). Any inference made based on a *p*-value, then, should be informed by the effectiveness with which your SCED was able to control for confounding variables that threaten the internal validity of any experiment (see Michiels & Onghena, 2018, for a discussion of the effects of linear trend and autocorrelation on RTs). An added benefit of randomization is that it is a statistical control for extraneous variables such as state of the organism at time *X* or state of the environment at time *X* (Edgington, 1984; Kratochwill & Levin, 2010; Onghena, 2018b). SCEDs and RTs are complementary, so any proposition based on the statistical inferences of RTs should also be based on SCED logic, and for that matter, sound behavioral theory.

Nonstatistical Inference

RTs make statistical inferences about that which you did as an experimenter and immediately observed as a result. RTs do not make

statistical inferences about the larger population, for they do not involve a random sample of that population. Due to the widespread misinterpretation of *p*-values, since before Cohen (1994) even described it, we must state and restate that which RT *p*-values can and cannot do. That RTs do not make statistical inferences about the larger population, however, does not mean that they are not relevant to a larger population of interest.

The General Process Approach

Schedules of reinforcement have been described as "the most powerful independent variables ever seen in psychology" (Zeiler, 1984, p. 485), not because of their magnitude of effect measured in terms of "effect size," but because of the orderly patterns of behavior they engender across species. When Skinner (1956) presented the cumulative records of a pigeon, rat, and monkey—each responding to a "multiple fixed-interval fixed-ratio schedule" of reinforcement—there was no telling the difference between their curves (p. 230). "Mice, cats, dogs, and human children could have added other curves to this figure" (Skinner, 1956, p. 231). Even with the addition of those other organisms, though, there would be no apparent difference in observed patterns of responding. Schedules of reinforcement exemplify what is meant by "general process approach," as behavior analysts study behavioral processes that cut across the Kingdom Animalia, and arguably, other phyla (e.g., it has been proposed that slime molds of the Kingdom Protoctista learn; see Boisseau, Vogel, & Dussutour, 2016).

The effects of reinforcement, punishment, extinction, discrimination training, and relational training on behavior are pertinent to organisms of all populations. Even where there are differences between populations, the differences are still attributable to behavioral processes to which any organism of any population is susceptible (e.g., gamblers exhibit a resistance to extinction, and if the conditions were right, people other than gamblers could exhibit that resistance too; see, e.g., Horsley, Osborne, Norman, & Wells, 2012). The generality of behavioral processes means that any experimental result emanating from the general process approach is about the larger population of interest (i.e., the population identified and defined based on your investigative goals). The use of RTs in pursuing your investigative goals means you do not need a random sample to make a statistical inference about the efficacy of your results. RTs allow you to say, "A difference in conditions caused a difference in behavior", while the general process approach allows you to extend that claim to other organisms by means of nonstatistical inference (Edgington & Onghena, 2007).

A cautionary note. Even if a general process approach is assumed, it does not guarantee an observed generality of effect. The burden of proof remains on the community of behavior analysts to show that arbitrarily applicable relational responding, for example, is a uniquely human and ubiquitous operant that accounts for a greater part of complex language and cognition (Hayes, Barnes-Holmes, & Roche, 2001). In order to elevate any finding to the level of generality and/or principle, behavior analysts must replicate across organisms, species, and contexts, as did Skinner and his contemporaries in the case of reinforcement schedules.

Critiques not Warranted

RTs are susceptible to the criticism that they do not make inferences about the larger population, but the same criticism can be charged against the parametric statistics that try and fail because they do not have a random sample. A large nonrandom sample does not equate to a stronger claim, and based on the replicability crisis, does not buy you generality of effect (Levenson, 2017). On this basis, RTs are not offered here as a mere alternative to parametric statistics; for example, when your distributional assumptions cannot be met. Instead, RTs are offered here as a replacement for the parametric tests based in a random sample model. There are conditions under which the random sample model might be feasible and warranted, but for the most part, the assumption of a random sample cannot be met. In his textbook, *Fundamental Statistics for the Behavioral Sciences*, Howell (2017) went so far as to say, "At the rate that the field is changing, I suspect that in the next 10 years randomization tests will largely replace not only the traditional nonparametric tests, but also the standard parametric tests such as *t* and *F*" (p. 525).

Conclusion

RTs make statistical inferences about what you did as an experimenter and immediately observed as a result. *P*-values are an index, so we must ask: "An index for what?" Skinner did not need an NHST *p*-value to know whether reinforcement was effective among other organisms of various populations, just as we did not need one to know that the earth is round (Cohen, 1994; 1995). By the same token, today we do not need parametric statistics to dictate the results of our behavior analytic studies, especially if we assume a general process approach: Behavior is lawful. The *p*-values of RTs complement the behavior analytic goal of predicting and controlling individual behavior because they are an index for what we did, as scientists and practitioners.

References

Baars, B. J. (1985). *The cognitive revolution in psychology*. New York: Guilford Press.

Baer, D. M., Wolf, M. M., & Risley, T. R. (1968). Some current dimensions of applied behavior analysis. *Journal of Applied Behavior Analysis, 1*(1), 91-97. https://doi.org/10.1901/jaba.1968.1-91

Barlow, D. H., & Hayes, S. C. (1979). Alternating treatments design: One strategy for comparing the effects of two treatments in a single subject. *Journal of Applied Behavior Analysis, 12*(2), 199-210. https://doi.org/10.1901/jaba.1979.12-199

Boisseau, R. P., Vogel, D., & Dussutour, A. (2016). Habituation in non-neural organisms: evidence from slime moulds. *Proceedings of the Royal Society B: Biological Sciences, 0160446*. https://doi.org/10.1098/rspb.2016.0446

Branch, M. N. (2018). The "reproducibility crisis:" Might the methods used frequently in behavior-analysis research help? *Perspectives on Behavior Science*, 1-13. https://doi.org/10.1007/s40614-018-0158-5

Bulté, I., & Onghena, P. (2008). An R package for single-case randomization tests. *Behavior Research Methods, 40*(2), 467-478. https://doi.org/10.3758/BRM.40.2.467

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49*(12), 997-1003. https://doi.org/10.1037/0003-066X.49.12.997

Cohen, J. (1995). The earth is round ($p < .05$): Rejoinder. *American Psychologist, 50*(12), 1103. https://doi.org/10.1037/0003-066X.50.12.1103

Dewey, J. (1941). Propositions, warranted assertibility, and truth. *The Journal of Philosophy, 38*(7), 169-186. https://doi.org/10.2307/2017978

Dugard, P. (2014). Randomization tests: A new gold standard? *Journal of Contextual Behavioral Science, 3*, 65-68. https://doi.org/10.1016/j.jcbs.2013.10.001

Edgington, E. S. (1966). Statistical inference and nonrandom samples. *Psychological Bulletin, 66*(6), 485-487. https://doi.org/10.1037/h0023916

Edgington, E. (1980). Random assignment and statistical tests for one-subject experiments. *Behavioral Assessment, 2*(1), 19-28.

Edgington, E. S. (1984). Statistics and single case analysis. In M. Hersen, R. M. Eisler, & P.M. Miller (Eds.), *Progress in behavior modification* (Vol. *16*, pp. 83-119). Orlando, FL: Academic Press, Inc.

Edgington, E. S. (1987). Randomized single-subject experiments and statistical tests. *Journal of Counseling Psychology, 34*(4), 437-442. https://doi.org/10.1037/0022-0167.34.4.437

Edgington, E., & Onghena, P. (2007). *Randomization tests* (4th ed.). Boca Raton, FL: Chapman & Hall/CRC.

Fisher, R. A. (1929). The statistical method in psychical research *Proceedings of the Society for Psychical Research, 39*, 189-192.

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics, 33*(5), 587-606. https://doi.org/10.1016/j.socec.2004.09.033

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, P values, confidence intervals, and power: A guide to misinterpretations. *European Journal of Epidemiology, 31*(4), 337-350. https://doi.org/10.1007/s10654-016-0149-3

Hayes, S. C. (1991). The limits of technological talk. *Journal of Applied BehaviorAnalysis, 24*(3), 417-420. https://doi.org/10.1901/jaba.1991.24-417

Hayes, S. C., Barlow, D. H., & Nelson-Gray, R. O. (1999). *The scientist practitioner: Research and accountability in the age of managed care* (2nd ed.). Needham Heights, MA: Allyn & Bacon.

Hayes, S. C., Barnes-Holmes, D., & Roche, B. (2001). *Relational frame theory: A post-Skinnerian account of human language and cognition*. New York: Springer Science & Business Media.

Heyvaert, M., & Onghena, P. (2014). Randomization tests for single-case experiments: State of the art, state of the science, and state of the application. *Journal of Contextual Behavioral Science, 3*(1), 51-64. https://doi.org/10.1016/j.jcbs.2013.10.002

Hofmann, S. G., & Hayes, S. C. (2018). The future of intervention science: Process-based therapy. *Clinical Psychological Science*, 1-14. https://doi.org/10.1177/2167702618772296

Holland, P. W. (2005). Counterfactual reasoning. In B. S. Everitt & D.C. Howell (Eds.), *Encyclopedia of statistics in behavioral science* (Vol. *1*, pp. 420–422). Chichester: John Wiley & Sons, Ltd.

Horsley, R. R., Osborne, M., Norman, C., & Wells, T. (2012). High-frequency gamblers show increased resistance to extinction following partial reinforcement. *Behavioural Brain Research, 229*(2), 438-442. https://doi.org/10.1016/j.bbr.2012.01.024

Howell, D. C. (2017). *Fundamental statistics for the behavioral sciences* (9th ed.). Boston, MA: Cengage Learning.

Huo, M., & Onghena, P. (2012). RT4Win: a windows-based program for randomization tests. *Psychologica Belgica, 52*(4), 387-406. https://doi.org/10.5334/pb-52-4-387

Kantor, J. R. (1953). *The logic of modern science*. Oxford, UK: Principia Press.

Killeen, P. R. (2018). Predict, control, and replicate to understand: How statistics can foster the fundamental goals of science. *Perspectives on Behavior Science*, 1-24. https://doi.org/10.1007/s40614-018-0171-8

Kratochwill, T. R., & Levin, J. R. (2010). Enhancing the scientific credibility of single-case intervention research: Randomization to the rescue. *Psychological Methods, 15* (2), 124-144. https://doi.org/10.1037/a0017736

Leek, J., McShane, B., Gelman, A., Colquhoun, D., Nuijten, M., & Goodman, S. (2017). Five ways to fix statistics. *Nature, 551*, 557-558. https://doi.org/10.1038/d41586-017-07522-z

Levenson, R. W. (2017). Do you believe the field of psychological science is headed in the right direction? *Perspectives on Psychological Science, 12*(4), 675-679. https://doi.org/10.1177/1745691617706507

Lilienfeld, S. O. (2017). Psychology's replication crisis and the grant culture: Righting the ship. *Perspectives on Psychological Science, 12*(4), 660-664. https://doi.org/10.1177/1745691616687745

Mace, F. C., & Critchfield, T. S. (2010). Translational research in behavior analysis: Historical traditions and imperative for the future. *Journal of the Experimental Analysis of Behavior, 93*(3), 293-312. https://doi.org/10.1901/jeab.2010.93-293

Michiels, B., & Onghena, P. (2018). Randomized single-case AB phase designs: Prospects and pitfalls. *Behavior Research Methods*, 1-23. https://doi.org/10.3758/s13428-018-1084-x

Moore, J. (1995). The foundations of radical behaviorism as a philosophy of science: A review of *Radical Behaviorism: The philosophy and the science* by M. Chiesa. *The Behavior Analyst, 18*(1), 187-194. https://doi.org/10.1007/BF03392706

Onghena, P. (2018a). Randomization tests or permutation tests? A historical and terminological clarification. In V. Berger (Ed.), *Randomization, masking, and allocation concealment* (pp. 209-227). Boca Raton, FL: Chapman & Hall/CRC Press.

Onghena, P. (2018b). Randomization and the randomization test: Two sides of the same coin. In V. Berger (Ed.), *Randomization, masking, and allocation concealment* (pp. 185-207). Boca Raton, FL: Chapman & Hall/CRC Press.

Onghena, P., & Edgington, E. S. (2005). Customization of pain treatments: Single-case design and analysis. *The Clinical Journal of Pain, 21*(1), 56-68. https://doi.org/10.1097/00002508-200501000-00007

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), aac4716. https://doi.org/10.1126/science.aac4716

Pashler, H., & Wagenmakers, E. J. (2012). Editors' introduction to the special section onreplicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science*, 7(6), 528-530. https://doi.org/10.1177/1745691612465253

Perone, M., & Hursh, D. E. (2013). Single-case experimental designs. In G. J. Madden, W.V. Dube, T. D. Hackenberg, G. P. Hanley, & K. A. Lattal (Eds.), *APA handbook of behavior analysis* (Vol. *1*, pp. 107-126). Washington, DC: American Psychological Association.

Pitman, E. J. G. (1937a). Significance tests which may be applied to samples from any populations. *Supplement to the Journal of the Royal Statistical Society, 4*(1), 119-130. https://doi.org/10.2307/2984124

Pitman, E. J. G. (1937b). Significance tests which may be applied to samples from any populations. II. The correlation coefficient test. *Supplement to the Journal of the Royal Statistical Society, 4*(2), 225-232. https://doi.org/10.2307/2983647

Pitman, E. J. G. (1938). Significance tests which may be applied to samples from any populations: III. The analysis of variance test. *Biometrika, 29*(3/4), 322-335. https://doi.org/10.2307/2332008

Putnam, H. (2002). Pragmatism and nonscientific knowledge. In U. M. Żegleń & J. Conant (Eds.), *Hilary Putnam: Pragmatism and realism* (pp. 14-24): London, UK: Routledge.

Putnam, R. A. (2002). Taking pragmatism seriously. In U. M. Żegleń & J. Conant (Eds.), *Hilary Putnam: Pragmatism and realism* (pp. 13-20): London, UK: Routledge.

Saltelli, A., & Stark, P. (2018). Fixing statistics is more than a technical issue. *Nature, 553*(7688), 281-281. https://doi.org/10.1038/d41586-018-00647-9

Shadish, W. R., Hedges, L. V., Pustejovsky, J. E., Boyajian, J. G., Sullivan, K. J., Andrade, A., & Barrientos, J. L. (2014). A d-statistic for single-case designs that is equivalent to the usual between-groups d-statistic. *Neuropsychological Rehabilitation, 24*(3-4), 528-553. https://doi.org/10.1080/09602011.2013.819021

Sidman, M. (1960). *Tactics of scientific research: Evaluating experimental data in psychology* (Vol. *5*). New York: Basic Books.

Skinner, B. F. (1956). A case history in scientific method. *American Psychologist, 11*(5), 221-233. https://doi.org/10.1037/h0047662

Skinner, B. F. (1971/1999). Why are the behavioral sciences not more effective? In V. G. Laties & A. C. Catania (Eds.), *Cumulative record: Definitive edition* (pp. 467-474). Acton, MA: Copley Publishing Group.

Stevenson, A., & Lindberg, C. A. (Eds.) (2015). *New Oxford American dictionary*. Digital Version: Oxford University Press.

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on *p*-values: Context, process, and purpose. *The American Statistician, 70*(2), 129-133. https://doi.org10.1080/00031305.2016.1154108

Zeiler, M. D. (1984). The sleeping giant: Reinforcement schedules. *Journal of the Experimental Analysis of Behavior, 42*(3), 485-493. https://doi.org/10.1901/jeab.1984.42-485

Zelen, M. (1998). Inference. In P. Armitage & T. Colton (Eds.), *Encyclopedia of biostatistics* (pp. 2035-2046). New York, NY: John Wiley & Sons, Ltd.

Received: August 12, 2018
Final Acceptance: January 13, 2019
Editor in Chief: Michael Young
Associate Editor: Todd McKerchar