American Journal of Criminal Justice
https://doi.org/10.1007/s12103-025-09845-4

To What Are We Inferring? The Widespread Misuse of Inferential Testing in the Most Cited Criminology and Criminal Justice Journals

J. Pete Blair[1,2] · Peter T. Tanksley[1] · Emily D. Spivey[3] · M. Hunter Martaindale[1,2]

Accepted: 2 July 2025

Abstract

Null Hypothesis Significance Testing (NHST) is widely used in criminology and criminal justice journals. Where either random selection of a sample or random assignment to treatment and control groups is absent, the meaning of inferential testing becomes unclear. Using a sample of articles published in 18 of the top ranked (h5-index) journals in criminology, criminal law, and policing, we examine (1) how often authors use NHST after violating the assumption of random selection/assignment, (2) whether authors focus more on statistical rather than substantive importance, and (3) whether this leads authors to ignore large statistically non-significant effects based on a flawed test. Our results suggest that it is commonplace to apply statistical procedures and interpret results with little attention to how data were generated and how results should be assessed. We recommend the use of exploratory research methods, better statistical training for students, and addressing publishing standards in our field.

Keywords Statistical significance · Random selection · Random assignment · Null hypothesis significance testing · Criminology · Exploratory data analysis

✉ Peter T. Tanksley
peter_tanksley@txtstate.edu

[1] Advanced Law Enforcement Rapid Response Training (ALERRT) Center, Texas State University, San Marcos, TX, USA

[2] School of Criminal Justice, Texas State University, San Marcos, TX, USA

[3] Criminal Justice Department, New Mexico State University, Las Cruses, NM, USA

Published online: 30 July 2025

Introduction

Inferential statistical tests such as null hypothesis significance testing (NHST), have been described as a mainstay of social sciences for empirically testing relationships and establishing the importance of empirical results (Bushway et al., 2006). NHST is widely taught in criminology and criminal justice academic programs and widely used in journals in these fields. Yet, NHST has been criticized virtually from its inception (Berkson, 1938; Blair, 2018; Boring, 1919; Carver, 1993; Cohen, 1994; Gigerenzer, 1987; Gill, 1999; Killeen, 2005; Kish, 2017; Kringen et al., 2018; Levine et al., 2008; Morrison & Henkel, 2006; Nickerson, 2000; Rouder et al., 2009; Rozeboom, 1960). Among the criticisms of NHST are the lack of a logically coherent framework, the meaninglessness of disproving a point estimate of 0, the arbitrariness of the standard alpha level (i.e., $p < 0.05$), the use of a hard decision rule, *p*-values being a function of both sample and effect sizes, and the widespread misunderstanding of *p*-values (Szucs & Ioannidis, 2017).

More recently, scholars have pointed to the cumulative effect of these issues, leading to concerns regarding confidence in the body of scientific evidence. This is perhaps most evident in the "replication crisis" observed in psychology (Open Science Collaboration, 2015; Szucs & Ioannidis, 2017). Research suggests the fields of criminology and criminal justice are not immune to these concerns, considering the lack of replication studies in criminology (Pridemore et al., 2018), evidence of publication bias (Crow et al., 2023), and the large portion of studies with low levels of statistical power in top criminology and criminal justice journals (Barnes et al., 2020). Another potential concern, we argue, is the misapplication of NHST. Although researchers have examined the misuse of NHST via questionable research practices, such as *p*-hacking and "HARKing" (i.e., Hypothesizing After Results are Known) (Chin et al., 2023), little attention has been paid to the extent of the misapplication of NHST within criminology and criminal justice.

One exception is found in Bushway and colleagues' (2006) attempt to assess whether research within criminology and criminal justice follows best practices for conducting NHST. The authors examined the flagship journals *Criminology* and *Justice Quarterly* as well as a review of experiments in criminology to assess whether the field appeared to be using statistical best practices when conducting NHST. Their review was focused on four primary issues: (1) the availability of the information necessary to determine the size of an effect; (2) the interpretation of results informed by the size of the effect (i.e., not just statistical significance); (3) correct handling of statistically non-significant results; and (4) the avoidance of basic errors in the application of statistical significance testing.

Bushway et al.'s (2006) analysis suggested that articles did generally present the information needed to determine effect sizes and avoided basic errors in statistical significance testing. Nevertheless, they often did not substantively interpret effect sizes and often did not handle statistically non-significant results well. The authors were particularly concerned with overreliance on binary accept/reject

decisions in the interpretation of results and the lack of substantive interpretation of the size of the reported effects. Bushway et al. argued that researchers seemed to be focused on statistical instead of substantive significance. This makes connecting the results in question to other research difficult, and statistical significance alone is not a sound foundation upon which policy should be based. Their results suggest that the core practice of inferential testing in the field is soundly conducted (at least in some of the top journals), although there is room for substantial improvement in how criminology and criminal justice researchers report and treat inferential statistical tests.

Researchers in criminology and criminal justice may regularly fail to meet the most basic assumption required to invoke inferential testing: randomization. Where either random selection of a sample or random assignment to treatment and control groups is absent, the meaning of inferential statical testing becomes unclear, and the dismissal of variables as unimportant based on statistical significance is not justified. When this is coupled with Bushway et al.'s (2006) finding that researchers tend to focus on statistical as opposed to substantive significance, it is possible that much of our understanding of crime and the response to crime is distorted by inappropriately applied NHST. Given criminology and criminal justice are applied fields, in which policy and practice are increasingly "evidence-based," distortions in our understanding can lead to ineffective (Lattimore et al., 2016), or even harmful policies, services, and interventions (Zane et al., 2016).

Inferential Testing

At its core, inferential testing attempts to guess from what we have observed to something that we have not observed. This jump from what we observed to something else requires information that goes beyond the data. Specifically, the appropriate use of inferential testing requires information about how the data were generated (Berk, 2004).

In the social sciences, there are two common situations under which inferential testing is done. The first is guessing what a population looks like from a sample. The second is guessing the values that we are likely to observe if we were to repeat an experiment many times. The first assumption necessary to invoke inferential testing is slightly different in each of these cases. In the case of guessing from a sample to a population, the assumption is that the sample was randomly drawn from the population. In the case of guessing from an observed experiment to the range of values that are likely to be obtained if the experiment were repeated many times, the assumption is that participants were randomly assigned to conditions. For the purposes of this paper, we will combine these different assumptions and refer to the new construct as the assumption of random selection/assignment. Although these assumptions serve distinct purposes, we justify their joint treatment in the following discussion and analysis, as both represent essential first steps in observational and experimental designs that aim to infer findings beyond the sample or setting.

When conducting an inferential test, the researcher takes a random sample from the population (or randomly assigns participants to conditions) and observes

the value(s) of interest. These are usually some sort of central tendency and variance. This information, coupled with the size of the sample, is used to estimate a confidence interval that represents the range of values that are likely within the population (or numerous replications of the experiment). These are the values that are being inferred.

In the NHST framework, if the confidence interval contains zero (indicating no relationship) the conclusion is that we cannot be confident that a relationship exists in the population (or that the result observed in the experiment was not the product of chance). In many non-significant cases, it is possible that the observed relationship is in the opposite direction of the relationship in the population (or repeated experiments) (Gelman & Carlin, 2014; Gelman et al., 2020).

The probability theory that underlies inferential tests and the creation of confidence intervals rests on the assumption of random selection/assignment. As Berk (2004) notes, when the assumption of random selection/assignment is not met, to what we are inferring is unclear. It is difficult to argue that the confidence intervals that are the basis of NHST represent an expected range of values of the population (or experimental replications) when the most basic assumption of their creation is not met. If the meaning of the test is unclear, basing judgments on that meaning would appear to be unjustified.

Ultimately, many other assumptions must be met for inferential testing to function appropriately (see Berk, 2004 for an in-depth discussion of these assumptions and the consequences of violating them), but for the purposes of this paper, we limit ourselves to assessing just the presence or absence of the random selection/assignment assumption because if this assumption is not met, statistical significance testing for the purpose of inference is generally not justified. We will not attempt to address whether the samples were actually randomly selected or representative of the population that is purportedly being assessed (see Berk, 2004 for a discussion of some common issues). We simply assess if researchers claim random selection/assignment.

It is important to note that arguments exist for the use of NHST in the absence of random sampling (Berk, 2004; Gelman et al., 2020; Kass, 2011). For instance, Gelman and colleagues (2020) argue that it is acceptable to use inferential testing on a non-random sample *if* one acknowledges the weaknesses of the data and makes specific adjustments to the sample to account for "certain known differences between sample and population" (p. 56). In addition to having a priori knowledge about a population, these adjustments present a host of other potential difficulties (see Berk, 2004 for a discussion). Similar logic is used in defense of NHST in the absence of random assignment: adjustments must be made for differences in treatment and control groups (Gelman et al., 2020).

It is also worth noting Gelman and colleagues (2020) have argued that inference can be used on population data (e.g., data from consecutive elections, data on all 50 states) if one thinks of this observed data and the associated outcome as "a random sample from a hypothetical superpopulation" (p. 154). Stated more explicitly, the errors are treated as "a random sample from the normal error distribution" (Gelman et al., 2020, p. 154). According to Gelman et al., these methods

are justified if one is interested in the hypothetical outcome in some future year (see also Redelings et al., 2012).

Setting aside the veracity of the above arguments, the defense of NHST in these contexts must be explicitly stated by scholars who wish to utilize them, and, in some cases, adjustments must be made to the data. For instance, Gelman et al.'s arguments for using NHST in the absence of random selection/assignment only apply if a researcher (1) acknowledges the weaknesses of their data and (2) takes the necessary steps to adequately adjust the data to account for differences between the sample and the population (or between the treatment and control groups) to the extent necessary. If these steps are not taken, researchers cannot invoke these defenses. Regarding population data, researchers must acknowledge that they are interested in making predictions about a hypothetical outcome in the future. That is, if researchers simply use NHST on non-random samples or population data and report results as if they had a simple random sample, the arguments made by Gelman et al. and others do not apply. This is the misapplication of NHST that we are specifically interested in assessing.

Because, to our knowledge, there has not been an attempt to identify how often inferential testing is conducted without random selection/assignment in criminology and criminal justice, we seek to answer the following research question:

RQ 1: How common is inferential testing without random selection/assignment in criminology and criminal justice journals?

Common Violations of the Random Selection/Assignment Assumption

Our experience suggests two common situations where inferential testing is conducted after violating the assumption of random selection/assignment. In the first, researchers use a sample collected using a non-probability sampling method (e.g., convenience sampling, purposive sampling, quota sampling). The second involves the use of population data.

Non-probability Samples A non-probability sample is best understood in relation to its opposite, a probability-based sample. For a sample to be considered a probability-based sample, the method used for its construction must meet the following requirements: (1) all cases are selected randomly from the population and (2) the probability of being selected is known (Singleton & Straits, 2010). In contrast, non-probability samples consist of cases for whom the probability of selection is unknown due to the lack of random selection (e.g., convenience sampling, quota sampling, purposive sampling). Consequently, the population to which any result might be generalized is also unknown for non-probability samples meaning that "the laws of probability do not apply" (Singleton & Straits, 2010, p. 158). Characterization of non-probability samples using descriptive statistics is appropriate and easy to understand. Inferential tests in such samples are inherently difficult (or inappropriate) to interpret given that the population to which one is inferring is undefined. The experimental equivalent of non-probability samples are quasi-experimental

designs. While there is substantial variation in these designs, they often involve non-random assignment to conditions. These designs, therefore, violate the assumption of random assignment. As with non-probability samples, the meaning of descriptive analyses comparing the groups is still straightforward, but the meaning of inferential testing is again unclear.

Sometimes when researchers violate the assumption of random selection/assignment, they make an argument about why inferential testing is still valid. There are different forms that this can take (including some very sophisticated statistical arguments that no longer infer to a population but instead to a model), but each of them involves bringing substantial information from outside of the data to argue that inference is still valid (see Berk, 2004, chapter 4 for a thorough discussion of these arguments and the problems with these approaches; see Gelman et al., 2020 for an example of adjustments made to allow inferences with a non-random sample). In the quasi-experimental case, sophisticated statistical measures (e.g., propensity score matching) might be used to establish treatment and control groups that are statistically equivalent. A researcher might also use statistical adjustments in the results to try and control for differences in groups (e.g., multiple regression or difference-in-differences).

While it is positive in these cases that the researchers at least recognize that they have violated the assumption of random selection/assignment and try to correct for this issue, it is not clear that these solutions allow for valid inference. In the examples above, the potential threat posed by unobserved variables looms large.[1] In the following analysis, we will evaluate articles on whether the authors met the assumption of random selection/assignment and, if they did not, if they justified their use of NHST. When encountered, we did not attempt to assess the validity of such justifications. We simply noted that random selection/assignment was not met, and that the researcher made some form of argument about why inference was justified.

Population Data The second situation in which NHST is used after violating the assumption of random selection/assignment occurs when the researcher uses a sample that is, in fact, not a sample but a population. This occurs when data on the entire entity one might wish to make inferences about is available (e.g., all officers in a police department, all traffic stops in a city). In this instance, NHST is unwarranted as any observed differences within the data are "real differences" in the population. The application of NHST to a population is inappropriate because (1) there is not a larger population that the results of NHST would apply to and (2) using NHST would likely give extra weight to findings that is unwarranted and induce the reader to assume the findings apply to different populations (which they do not).

A similar situation arises when two sets of population-level data are compared using NHST to assess if they are different from one another. What such a test

[1] We are not arguing that these methods are invalid or undesirable, only that it is not clear that inference to some larger population (or repetitions of the experiment) are justified.

might reveal is also unclear. Observed differences between two populations are real differences. A statistically significant test coefficient should not be needed (or taken) as evidence that any differences between populations are more likely to be real—they are real. Contrariwise, any statistically non-significant test coefficient comparing two populations should not be taken as evidence that observed differences between populations are less likely to be real—they are *still* real. At the level of populations, inferential tests usually only serve to muddy the interpretation of otherwise straightforward descriptive analyses.

We must again acknowledge the arguments of Gelman et al. and others for the use of inferential testing on population data for the purposes of making predictions about hypothetical outcomes in the future. Whether or not one accepts this argument as valid, we fear it may lead to scholarly "hand waiving" in which authors cite the arguments as more or less an endorsement to use inferential testing on population data in any circumstance.

No assessment to date has examined these situations, so we address these issues as research questions:

RQ 2*: How often is inferential testing used with non-probability samples or populations?*
RQ 3*: When an inferential test is used with non-probability sample or a population, how often is a justification not provided for why this was done?*

The Impact on Interpretation

It may be that this inappropriate use of inferential testing is harmless, but as Bushway et al. (2006) noted, researchers often appear to rely on statical significance testing to determine substantive significance. Results that are found to be statistically significant are considered substantively important but results that are not statistically significant are not (or in some cases treated as zero effect). As Bushway et al. note, statistical significance is not the same as substantive significance. When the assumption of random selection/assignment is not met, it is difficult to argue that any importance should be placed upon statistical significance. If statistical significance is used as the determinant of substantive significance when the test is invalidly applied, this could mean that substantively important effects are being ignored. If this occurs on a large scale, it is possible that large areas of our knowledge are substantially distorted by the misuse of inferential testing. Based upon Bushway et al.'s results, we expect to commonly find that statistical significance is substituted for substantive importance, but this issue has not been explored in conjunction with violating the assumption of random selection/assignment, so we address this issue as a research question:

RQ 4: When the assumption of random selection/assignment is violated, how often are results treated as substantively important based solely/principally on statistical significance?

A corollary of equating statistical significance with substantive importance is the issue of ignoring statistically non-significant effects altogether, even if they are larger than the observed statistically significant ones. Because statistical significance is impacted by variance, it is possible to have a statistically non-significant effect that is larger in observed magnitude than a statically significant effect. When inferential testing is performed correctly, discounting of the statistically non-significant effect may be justified because we cannot be confident in the direction of the effect, but when the assumption of random selection/assignment is not met, it is difficult to justify ignoring a statistically non-significant effect that is large. When inferential testing is inappropriately used to determine whether a variable is substantively meaningful, and this results in ignoring effects that are larger than the statistically significant effects, researchers may be painting an inaccurate picture of the world. Researchers may be ignoring large and substantively important effects simply because they fail to reach statistical significance in an inappropriately applied test. This leads to our fifth and final research question:

RQ 5: When the assumption of random selection/assignment is violated, how often do researchers ignore statistically non-significant effects that are as large or larger than the statistically significant effects?

While we have stated these issues as research questions, we suspect the following are occurring frequently in the field: (1) researchers are using inferential tests when they have violated the core assumption on which these tests are premised; (2) the researchers then use these invalid tests to judge whether a variable has substantive importance. If we are correct, the implication is that our knowledge of crime and the criminal justice system may be substantially distorted by the misapplication of inferential tests.

Methodology

Sample

We collected a sample of the 20 most cited journals in criminology and criminal justice. We used rankings provided by Google Scholar of the top 20 journals in a variety of fields based on their h5-index. The h5-index is defined as "the largest number h such that at least h articles in that publication were cited at least h times each" for only those articles published in the last five years (Google Scholar, 2024). We chose the social science subcategory of "criminology, criminal law, and policing," which were posted May of 2021 (Table 1).

For each journal, we selected the most recent regular, complete issue of that journal and then randomly selected 5 articles in that issue for a total initial sample of 100 articles. We believe that this stratified random sampling strategy provided us with an accurate snapshot of the most cited journals in criminology and criminal justice and allowed us to keep coding at the article level instead of having to aggregate results by journal because of the different numbers of articles in each journal issue.

Table 1 The top 20 criminology and criminal justice journals and their contribution to the analytic sample

Rank[1]	Name	Volume Number	Issue Number	Articles in Issue
1	British Journal of Criminology	61	3	14
2	Journal of Criminal Justice	73	-	17
3	Criminal Justice & Behavior	48	6	8
4	Crime & Delinquency[2]	67	6, 7	11
5	Law & Human Behavior	45	1	5
6	Justice Quarterly	38	4	7
7	Criminology & Public Policy	20	1	6
8	Journal of Quantitative Criminology	37	1	10
9	Criminology	59	1	6
10	Journal of Experimental Criminology	17	1	8
11	Journal of Research in Crime & Delinquency[3]	58	3, 4	7
12	Sexual Abuse	33	4	5
13	European Journal of Criminology	18	3	8
14	International Journal of Offender Therapy & Comparative Criminology	65	8	8
15	Policing & Society	31	4	7
16	Theoretical Criminology	25	2	8
17	Criminology & Criminal Justice	21	2	7
18	Criminal Justice Policy Review	32	5	6
19	Youth Violence & Juvenile Justice	19	2	5
20	American Journal of Criminal Justice	46	2	9

[1]Ranked according to the h5-index as of May 2021

[2]The selected issue from the journal *Crime & Delinquency* was actually a combination of two issues (#6 and #7)

[3]Issue 4 of Volume 58 in the *Journal of Research in Crime and Delinquency* only contained three articles. To account for this, we combined issues 3 and 4 for a total of seven articles, from which we randomly selected five articles

The number of articles in each issue is often dependent on the number of issues released annually. Most of the journals included in the sample release issues either bi-monthly ($n=8$, 40%) or quarterly ($n=8$, 40%). The number of articles included in each journal in the sample ranged from 5 to 17 with a mean of 8.1 articles. In the issues limited to five articles, all articles in the issue were selected for inclusion in the sample.

Article Coding

For each article, we applied decision rules that allowed us to dichotomize articles in a manner that coincided with our research questions (Table 2). This approach allowed us to progressively categorize articles according to how they met/departed

Table 2 Research questions and the corresponding coding used for articles using inferential statistics ($n=50$)*

Research Question	Article Coding	
	Criteria	Score
1 How common is inferential testing without random selection/assignment in criminology and criminal justice journals?	Author(s) uses NHST without random selection/assignment	Yes = 1 No = 0
2 How often is inferential testing used with non-probability samples or populations?	Author(s) uses NHST with a sample collected using a non-probability sampling method or with a population	Yes = 1 No = 0
3 When an inferential test is used with non-probability sample or a population, how often is a justification not provided for why this was done?	Author(s) failed to provide justification for use of NHST with a non-probability sample or with a population	Yes = 1 No = 0
4 When the assumption of random selection/assignment is violated, how often are results treated as substantively important based solely/principally on statistical significance?	Author(s) interpreted results solely or principally based statistical significance	Yes = 1 No = 0
5 When the assumption of random selection/assignment is violated, how often do researchers ignore statistically non-significant effects that are as large or larger than the statistically significant effects?	*For articles that used NHST in non-probability samples or populations and did not provide a justification*: Author(s) discounted/ignored results that were not statistically significant despite being large or larger than statistically significant effects in the model	Yes = 1 No = 0

*Some articles used inferential statistics but were excluded from further consideration due to the nature of the method used. Purely methodological articles and those using Bayesian or time-series methods are examples. NHST = null hypothesis significance testing

from proper use of inferential statistics under the assumption of random selection/assignment. For research questions 1–3, simple decision rules allowed for the dichotomization of articles according to whether they violated the assumption of random selection/assignment (RQ 1), used non-probability samples or populations (RQ 2), or provided justifications for their use of NHST (RQ 3).

The decision rule used to dichotomize articles based on the sole reliance of statistical significance (RQ 4) was adapted from Bushway et al. (2006) and included two criteria. The first criterion dealt with the statistical significance of the model. We examined if authors, after reporting the statistical significance of a variable, went on to describe their findings in the context of the overall model using metrics such as explained variance (i.e., R^2). The second criterion dealt with statistical significance of individual variables. We examined if authors used statistical significance as the primary method for evaluating the importance of variables in the discussion/conclusion section of the paper. In step with Bushway et al., this variable captured whether authors evaluated the importance of key independent variables outside of statistical significance. If an article did not meet either criterion, we coded the article as solely/principally relying on statistical significance.

Lastly, we examined if articles using non-probability samples or populations treated large statistically non-significant effects as substantively unimportant (RQ 5). We made this determination by converting reported coefficients into elasticities (% change in Y divided by % change in X) or using reported effect sizes. We compared the statistically significant coefficients with the statistically non-significant ones. If the elasticity or effect size of the smallest statistically significant coefficient was smaller than the elasticity or effect size of the largest statistically non-significant coefficient, and the paper discounted the substantive importance of the large statistically non-significant variable, then we coded the article as meeting this decision rule.

Reliability

One coder coded all 100 articles. A second coder coded a randomly selected subset of 20 articles. The intraclass correlation coefficient of the coders was 0.881, which is considered good reliability ($0.75 > ICC > 0.90$) (Koo & Li, 2016), and Krippendorff's alpha was 0.79, indicating acceptable reliability ($\alpha > 0.60$) (De Swert, 2012). Where there were disagreements, the lead author on the paper discussed the disagreement with the coders and the group arrived at a consensus.

Results

How Prevalent was NHST in Criminal Justice/Criminology?

Since Bushway et al.'s (2006) initial investigation, little attention has been paid to the misapplication of NHST in the fields of criminology and criminal justice. Specifically, it is unclear the extent to which criminologists and criminal justice researchers

are conducting inferential tests when violating the basic assumption upon which these tests are based. Of the 100 randomly selected articles from the field's top 20 journals, 63 articles used inferential testing. The types of articles excluded from the sample include qualitative pieces ($n=20$), theoretical pieces ($n=13$), or strictly methodological pieces ($n=4$). Further, several inferential articles were excluded from the final sample because they used time series analysis ($n=5$), Bayesian statistics ($n=1$), meta-analysis ($n=3$), or were strictly methodological pieces ($n=4$) (e.g., using inferential testing on model-generated simulation data). Two issues of the journals included in the sample did not include any inferential articles (*Theoretical Criminology,* volume 25, issue 2; *Criminology & Criminal Justice,* volume 21, issue 2). Thus, the final sample was comprised of 50 articles from 18 most-cited journals in criminology, criminal law, and policing.

How Often Was NHST Used Without Random Assignment or Selection?

In the 18 most-cited journals in the field, nearly three-quarters of articles that used inferential testing did not meet the assumption of random selection/assignment. In fact, only 13 articles (26%) met the assumption of random selection/assignment (Fig. 1, Panel A). Among these 13 articles, seven used some method of probability sampling to randomly select units from the population of interest, five randomly assigned units to treatment or control conditions, and one article employed both methods. These results suggest that using inferential tests without meeting the requirement of random selection/assignment is common in the most-cited criminology/criminal justice journals.

How Often Was NHST Used with Non-Probability Samples or Populations?

Of the 37 articles that inappropriately used inferential tests, 24 used a non-probability sample. Thirteen used inferential testing on population data. Again, we acknowledge Gelman et al.'s (2020) arguments regarding superpopulations. If authors used population data and cited Gelman et al.'s arguments for doing so, this was coded as a justification for inference. Notably, no articles using inferential testing on population data provided any such justification.

Was Justification Provided for the Misapplication of NHST?

Of the 37 articles that used inferential testing without random selection/assignment, 31 did not offer any justification for this practice. Among the six articles that attempted to justify this practice, five provided justifications for using data that were not randomly assigned, and one provided a justification for using data that were not randomly selected. Regarding the content of justifications, authors cited the use of quasi-experimental methods, such as matching procedures. More specifically, two studies used covariate exact matching, whereas three studies used propensity score matching. Generally, the justifications for using these methods note random assignment is not feasible in certain contexts and these

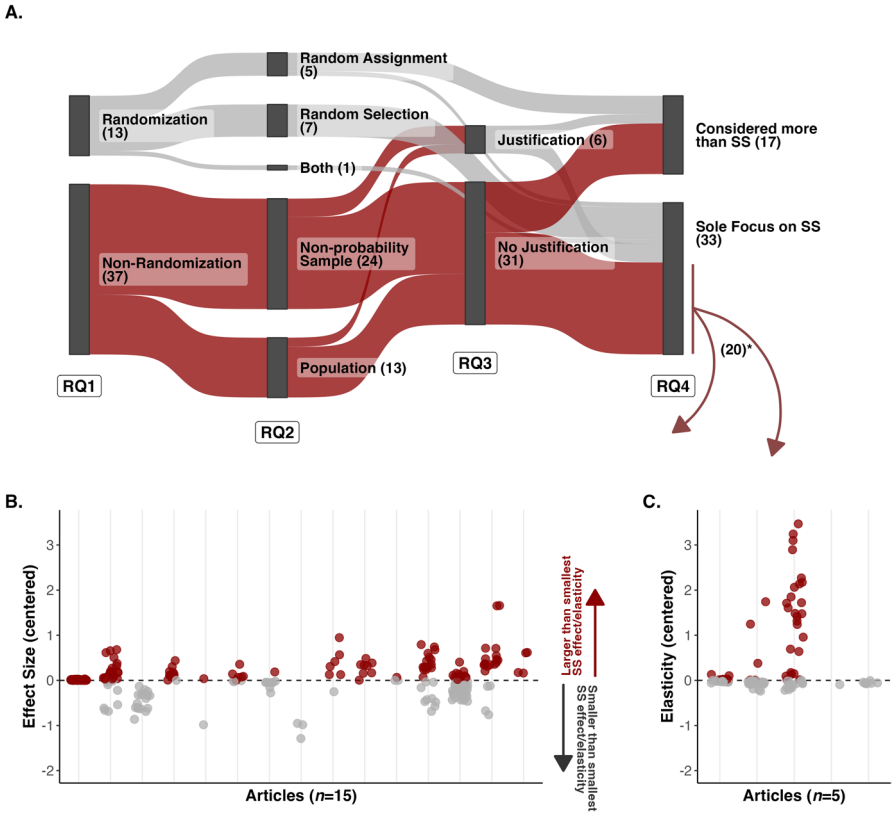

Fig. 1 Sankey plot (**Panel A**) showing the flow of included articles ($n = 50$) across research questions 1–4. Flows representing articles that did not meet the NHST assumption of random selection/assignment are filled in red; flows representing articles that did meet the assumption or provided a justification for not meeting it are filled in grey. Numbers in parentheses correspond to the number of articles at each stage. Research question 5 was examined using effect sizes (**Panel B**) and elasticities (**Panel C**) that were extracted/calculated using information from articles that did not (1) meet the random selection/assignment assumption, (2) provided no justification for continued use of NHST, and (3) relied solely/principally on statistical significance ($n = 20$). The points on **Panels B** and **C** correspond to statistically non-significant coefficients reported in analytic models of each article. Effect sizes were typically reported as odds ratios or incidence rate ratios. Elasticities were calculated when articles reported only unstandardized regression coefficients from linear models. Effect sizes/elasticities were centered by taking the absolute value of all coefficients in a model and then subtracting the smallest statistically significant coefficient for that model (all models from a given article are plotted on the same vertical line). We took the natural log of effect sizes prior to taking the absolute value to ensure sensible comparisons of positive and negative coefficients. Abbreviations: NHST = null hypothesis significance testing; RQ = research question; SS = statistically significant

quasi-experimental methods "mimic random assignment and balance covariates across treatment and control groups" (Levchak, 2021, p. 2). In terms of justifications for the use of inferential testing on data that were not randomly selected, one article cited several factors, including costs for travel for the research teams and other practical limitations. Overwhelmingly, though, authors did not provide

an argument for using inferential testing when the assumption of random selection/assignment was violated.

How Often did Researchers Equate Statistically Significance and Substantive Importance?

Of the 31 articles that did not use random selection/assignment nor offer a justification, 24 appeared to use statistical significance as the sole measure of substantive importance. Specifically, five articles (16%) did not consider the importance of the variables outside of statistical significance (e.g., explained variance), 13 articles (42%) used statistical significance as the principal means for assessing the importance of key variables, and six articles (19%) engaged in both practices. Of the 19 articles that *did* use random selection/assignment or offered a justification for not using it, 13 (68%) appeared to use statistical significance as the sole measure of substantive importance. Specifically, eight articles (42%) did not consider the importance of the variables outside of statistical significance, two articles used significance as the principal means for assessing the importance of key variables (11%), and three articles (16%) had evidence of both. Consistent with Bushway et al.'s (2006) findings, it appears that researchers still place a premium on the statistical significance of variables when assessing substantive importance.

How Often Did Researchers Ignore Statistically Non-Significant Effects that Were as Large or Larger than the Reported Statistically Significant Effects?

We extracted information from reported results of each article to produce effect sizes (e.g., odds ratios, incidence rate ratios) or elasticities, as well as their reported statistical significance. Next, we (1) compared the effect size/elasticity of the independent variable with the smallest statistically significant coefficient to the largest statistically non-significant coefficient and (2) determined whether the paper discounted the substantive importance of a larger statistically non-significant effect.

Twenty-four articles both violated the assumption of random selection/assignment (without including any justification) and used statistical significance as the sole measure of substantive importance. From these, we were able to obtain effect sizes (or calculate elasticities) for 20 articles. Four articles were not included because one article did not report any multivariate results, one did not have any statistically non-significant results, another did not report summary statistics (preventing the calculation of elasticities), and the final article reported results from a structural equation model with latent variables that do not lend themselves to the calculation of elasticities. We removed these four articles from this portion of the analysis. We additionally did not consider individual models within articles that contained multiplicative interaction terms as these shift the interpretation of coefficients in the model.

Of the 20 remaining articles, 16 (80%) reported statistically non-significant effect sizes/elasticities that were as large or larger than the smallest statistically significant effects in the model (Fig. 1, Panels B & C), and 12 (60%) of these articles appeared to discount these large statistically non-significant effects.

Discussion

Our review of the most cited journals in criminology, criminal law, and policing according to Google Scholar (as of May 2021) found that almost three-fourths of these articles used null hypothesis significance tests (NHST) after violating their most basic assumption: the assumption of random selection/assignment. Further, most authors of these articles (~75%) used NHST as the primary determinant of substantive importance. Almost the same proportion of articles that did not violate the assumption of random selection/assignment also appeared to use statistical significance as the primary arbiter of substantive significance. Additionally, more than half of the articles that violated the assumption and relied primarily on statistical significance also contained non-significant effects that were *larger* in effect size than at least some of the statistically significant effects. That is, these articles ignored observed effects that were larger than effects that were statistically significant based on a flawed test.

Given the nature of our stratified random sample of published journal articles, we could have reported 95 percent confidence intervals and conducted significance tests on all reported proportions. Had we done so, it would have begged the question: *to what are we inferring*? Some readers might assume that these confidence intervals would reflect the range of likely values for the entire population of criminology and criminal justice journals. In fact, such inferences would only apply to the remaining articles in the journal issues that were *not* included in our sample. Yet, the addition of such statistics would likely have had the effect of adding extra weight to our results. Indeed, the addition of statistical tests to the above descriptive analysis would have added almost no additional information because most of the articles from most of the journal issues were initially included in our sample. As we have done here, we advocate for researchers to avoid using NHST like the proverbial drunkard uses a lamppost: "for support rather than for illumination" (Lang, as cited in Lowndes, 1970, p. 6).

Overall, the findings of the current paper suggest our field appears to be engaged in the widespread "…ritualization of null hypothesis significance testing" (Cohen, 1994, p. 997). Often, we appear to be applying statistical procedures and interpreting the results mechanically with little attention to how data were generated and how results should be assessed. The implications are far from inconsequential, as these practices may have created substantial distortions in our understanding of, and responses to, crime.

Entrenchment of NHST in Criminal Justice/Criminology

To be sure, some scholars are doing thoughtful and well applied analysis in our field. This does not seem to be the norm, however. As we noted in the introduction, NHST has been criticized almost since its inception. These criticisms have been raised specifically in criminology and criminal justice for nearly two decades, yet it appears that our field is still widely, and often inappropriately, engaged in NHST. Bushway

et al.'s (2006) suggestion that substantive significance should outweigh statistical significance appears to have been largely disregarded. Likewise, Berk's (2004) observation that regression is very useful as a descriptive tool, less useful for population inference, and very questionable for causal inference, appears to have been overlooked.

There are likely many reasons for the widespread misapplication of NHST in criminology and criminal justice, including social factors in research (Sweeten, 2020). For instance, some have suggested an academic culture of "publish or perish" may contribute to other issues observed in the field, such as the lack of replication studies in criminology and criminal justice journals (Lösel, 2018) and the rarity of null findings published in top journals (Crow et al., 2023). This culture may contribute to the misapplication of NHST. Specifically, the pressure to publish may facilitate the routinization of NHST. This ever-increasing pressure to produce a steady stream of publications lends itself well to "cookbook" statistics. At the very least, the pressure to produce is unlikely to encourage any challenges to the current status quo.

Addressing the Misuse of NHST in Criminology/Criminal Justice

If a simple solution existed to the misapplication of statistics in criminological research, the problem would not persist, and there would be no need for the current paper. Gelman et al. (2020) argued that we should forget about statistical significance altogether because the tests waste information, the models are noisy, our populations do not contain true zeros, and comparisons and effects vary by context. Given the difficulty we seem to have in moving away from NHST as the default for statistical analysis and the associated "star chasing," it may be that this extreme approach is ultimately warranted. Indeed, qualitative data gathered from different stakeholders in the scientific system revealed several perceived obstacles to moving away from a reliance on NHST (Stunt et al., 2021). The most central theme was "reactivity", in which individuals are "disinclined to take responsibility and await external signals and initiatives of others" (p. 13). Stunt et al. concluded that a paradigm shift in the short term is not realistic; however, through small steps, "it is possible to decrease the scientific community's strong dependence on NHST and *p*-values" (p.22).

Many recommendations have been put forth to address disciplinary overuse of NHST, including pre-registration of hypotheses and analysis parameters, publishing raw data, increasing statistical power, better training, and teaching alternative approaches (Szucs & Ioannidis, 2017). Like crime and every other "wicked" problem the social sciences attempt to study/ameliorate, the misapplication of NHST will undoubtedly require a complex, multimodal approach to shift the culture in the discipline that gives it purchase. Though each facet of such an approach is worthy of an in-depth treatment, we limit ourselves to describing three that we believe would substantively and positively contribute to a culture shift toward better (and more limited/appropriate) use of NHST.

Suggestion #1: Embrace the EDA-CDA continuum One reason for the misuse of NHST in our field is the misalignment of statistical tools and researchers' intentions. For instance, criminologists often do not have the data, evidence base, or theories mature enough to support the strictest application of statistical tools like NHST—tools that are designed, at their core, for causal inference based on probability theory. Despite being the most coveted of statistical claims in all of science, we seriously doubt that causal inference is what most criminologists attempt to achieve with the data and theories they have at hand. Instead, we believe that most criminological research contains some portion—perhaps a major portion—of exploration (i.e., it is not solely confirmatory). Far from a negative, exploratory research is exceedingly valuable for a discipline that is not what Fife and Rodgers (2022) refer to as a "mature" science (i.e., having widely accepted and codified theories that generate testable predictions). What is problematic, however, is when criminologists (1) consciously or unconsciously ignore the exploratory nature of their research, (2) misapply statistical tools like NHST, and (3) infer beyond what their data and theory can support.

One approach to addressing the problem of misaligned tools and intentions is to embrace the exploratory/confirmatory data analysis (EDA-CDA) continuum (Fife & Rodgers, 2022). Most criminologists are familiar with CDA because, as was shown in this paper, they regularly use tools designed for it (e.g., t-tests, ANOVA, multiple regression). What most criminologists do not appreciate about these tools, as was also shown in this paper, are the strict assumptions that underly their use (e.g., random selection/assignment). CDA, sometimes referred to as "strict" CDA, represents one pole of the EDA-CDA continuum and is rarely (faithfully) used in practice, except in "mature" scientific disciplines (e.g., physics)—even disciplines like genetics rarely engage with strict CDA. Rather, they engage in "rough" CDA, where assumptions are relaxed somewhat, and inferences are (or should be) severely curtailed. A prime example of rough CDA is the process of model fitting for the latent variable models common in psychology. This iterative process includes specification, identification, estimation, re-specification, and replication (Bollen, 1989). By listening to the data, researchers are engaging in some degree of exploration and thus depart (however minutely) from strict CDA. Inferences may still be made based on rough CDA; they are more limited and local to the data used, but they provide support to their underlying theories and illuminate the way for future research (perhaps using strict CDA).

In contrast to rough CDA, criminologists rarely knowingly engage in the most basic form of research, that of EDA. EDA has existed for well over a century (Rodgers, 2010), though it was first codified into a philosophy/group of principles by Tukey in the 60's and 70's (e.g., Tukey, 1977). EDA uses several tools (e.g., boxplots, histograms) though it is best thought of as a general approach and not a set of specific tools. The goal of EDA is to search one's data for patterns of interest while honestly reporting the results of each step in the journey. It takes a certain amount of academic humility to conduct EDA because it requires a researcher to admit that they either (a) do not have good theories or (b) that their theories are tired and in need of new insights.

EDA is a method for probing data for relationships of interest by "listening to one's data" (Fife & Rodgers, 2022, p. 6). Because discovery is at the core of EDA, those who employ it are obliged to also respect their data and avoid double-dipping by engaging in (rough) CDA after a relationship of interest has been identified. Relationships observed using the tools of EDA are of unknown reality until they are validated using tools designed for confirmation *in a different dataset* from the one in which they were originally discovered. The separation of discovery and validation datasets (common in machine learning contexts) is undoubtedly an added burden for researchers; without such separation, however, researchers run the risk of compromising yet another assumption of CDA.

Given the non-probabilistic nature of much of the data used in criminological research, as well as relative "immaturity" of the discipline, we believe that EDA—or at least an honest acceptance of the more limited inferences afforded by rough CDA—will be crucial to improving the state of criminological research. Updating graduate curricula and publishing standards will be important for ensuring a new appreciation for the EDA-CDA continuum. Our following suggestions focus on these areas.

Suggestion #2: Revise graduate statistics training to emphasize real data, EDA, and alternatives to NHST Graduate training in statistics would benefit from substantive changes to both content and pedagogy. Echoing the American Statistical Association's GAISE recommendations for undergraduate instruction (2016), graduate programs should engage students with real-world data. Unlike sanitized textbook examples, real data often include missingness, outliers, and violations of model assumptions—complexities that reflect the messiness of applied research. Early exposure to these challenges prepares students to think critically about data, rather than relying on mechanical or decision-tree approaches to analysis.

This shift naturally fosters the use of EDA, which promotes data familiarity, highlights assumption violations, and enhances model development. When students are taught to document and justify decisions made during the exploratory process (e.g., how outliers were handled), programs also instill habits of transparency and rigor in reporting. These same principles extend beyond the modeling stage: EDA provides essential tools for post-estimation diagnostics, including residual analysis, identification of influential cases, and assessment of model fit. Incorporating EDA more explicitly into graduate curricula can build the analytic flexibility and justification skills that are foundational to responsible inference.

In tandem with EDA, statistical instruction should expand to include modern alternatives to NHST. Despite its ubiquity, NHST is frequently misunderstood—students often misinterpret p-values and receive little guidance on the actual inferential scope of the test. Bayesian and likelihood-based approaches offer more coherent inferential frameworks by modeling uncertainty and incorporating prior knowledge. These approaches complement EDA especially well, as exploratory analyses can inform the specification of prior distributions and highlight model assumptions that merit adjustment.

Together, these reforms—emphasizing real data, formalizing EDA, promoting transparency, and integrating modern inferential tools—can improve students' statistical literacy and cultivate more thoughtful, flexible researchers.

Suggestion #3: Reform publishing standards to support methodological alignment and transparency Efforts to implement the EDA-CDA continuum and improve graduate training are important steps forward. However, changes to the publication process are also necessary if the field is to shift away from the ritualized use of NHST. We offer two recommendations. First, editors and reviewers should require that manuscripts meet the assumptions necessary for inferential testing if authors wish to report inferential results. Journals can support this expectation by clearly articulating these requirements in their author guidelines and requiring authors to acknowledge them during the submission process.

When the assumptions of NHST are not met—random selection or assignment being only one example—authors should be encouraged to report descriptive findings or exploratory regressions (see Berk, 2004). In these cases, statistical significance testing should be avoided, as it is unclear to what these tests are inferring. Importantly, authors who choose to use NHST should be required to explicitly justify its use. For example, authors might acknowledge a violation of a model assumption but provide a defensible argument (e.g., via simulation or prior literature) that the violation does not substantively bias their results.

This approach encourages authors to argue for the relevance and importance of their findings based on context, design, and theory, rather than simply defaulting to statistical significance. In doing so, the field can begin to reframe many published studies for what they often are: context-bound case studies, offering meaningful insight within specific samples but not necessarily justifying population-level inferences. Such research is still valuable and should not be discounted by editorial staff solely for lacking a conventional inferential frame.

Our second recommendation pertains to the role of editors and reviewers in guiding disciplinary norms. Editors devote substantial effort to ensuring the theoretical and historical quality of manuscripts, but statistical rigor is often overlooked due to tradition or inertia. While editors face real pressures—ensuring a steady flow of publishable material, navigating the"publish or perish"environment—they also have a responsibility to address issues arising from the replication crisis and the misuse of inferential tests, as underscored in this paper.

We are not advocating for a rigid prohibition of NHST. Instead, we encourage editors to publicly support the submission and publication of EDA-focused research, particularly when CDA assumptions cannot be met. Doing so would reduce pressure to"star chase"and expand the types of rigorous empirical work journals are willing to publish. However, we also urge editors to extend this responsibility to their reviewer pool. Resistance to EDA-focused analyses often originates from reviewers steeped in an NHST-centric publishing culture. Editors must be prepared to mediate these conflicts, advocate for methodological appropriateness, and uphold standards that reflect both statistical integrity and disciplinary advancement.

Conclusion

Do our findings extend beyond the specific journals and issues that we selected? Our sample was undoubtedly representative of our specified population of journals/issues because we sampled almost all the articles it contained; whether our defined population was reflective of an even larger population of journals/issues (e.g., a super population of journals over time), however, is a judgement beyond the scope of this study. Despite this, we believe that the findings shared here would prove to be "typical" should our approach be replicated in the future with new samples. If anything, we suspect that the situation is more extreme in the less cited journals, although we have no way to examine this with our current sample.

It is possible that another study with a different sample might produce findings that conflict with those reported herein. If that occurs, our hope is that it will generate more discussion and facilitate more research on the extent of misapplication of statistical testing within criminology and criminal justice. Hopefully our efforts will be replicated to slowly chip away at the replication crisis in our field. Given the numerous critiques leveled against NHST, coupled with the findings of the current study, we recommend a shift away from NHST unless strict adherence to CDA principles is achieved. Ultimately, we hope that by no longer relying on tests that have unclear meanings, we will develop a deeper understanding of reality and better tools for addressing the causes and consequences of crime.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s12103-025-09845-4.

Funding The authors received no financial support for the current work.

Data Availability The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request. A list of the articles retained for descriptive analysis (N = 50) are available in the Supplemental Material.

Declarations

Conflicts of interest PTT is an editorial board member of the American Journal of Criminal Justice. This fact was disclosed to the editor during the initial submission and double-blind review ensured that peer reviewers were unaware throughout the review process.

References

Barnes, J. C., TenEyck, M. F., Pratt, T. C., & Cullen, F. T. (2020). How Power-Ful Is the Evidence in Criminology? On whether We Should Fear a Coming Crisis of Confidence. *Justice Quarterly, 37*(3), 383–409.

Berk, R. A. (2004). *Regression analysis: A constructive critique*. Sage Publications, Inc.

Berkson, J. (1938). Some difficulties of interpretation encountered in the application of the chi-square test. *Journal of the American Statistical Association, 33*(203), 526–536.

Blair, J. P. (2018). A consideration of significance testing. *ACSJ Today, 43*(1), 5–6.

Bollen, K. A. (1989). *Structural equations with latent variables*. John Wiley & Sons.

Boring, E. G. (1919). Mathematical vs. scientific significance. *Psychological Bulletin, 16*(10), 335.

Bushway, S. D., Sweeten, G., & Wilson, D. B. (2006). Size matters: Standard errors in the application of null hypothesis significance testing in criminology and criminal justice. *Journal of Experimental Criminology, 2*, 1–22.

Carver, R. P. (1993). The case against statistical significance testing, revisited. *The Journal of Experimental Education, 61*(4), 287–292.

Chin, J. M., Pickett, J. T., Vazire, S., & Holcombe, A. O. (2023). Questionable research practices and open science in quantitative criminology. *Journal of Quantitative Criminology, 39*(1), 21–51.

Cohen, J. (1994). The earth is round (p<. 05). *American psychologist*, *49*(12), 997.

Crow, M. S., Smykla, J. O., O'Brien, H., Cerna, T., Johnson, A., Pisaris, S.,…Wilder, J. (2023). What's in your file drawer? the case of the missing null in criminology and criminal justice. *Crime & Delinquency*, *69*(12), 2574–2594.

De Swert, K. (2012). Calculating inter-coder reliability in media content analysis using Krippendorff's Alpha. *Center for Politics and Communication, 15*(1–15), 3.

Fife, D. A., & Rodgers, J. L. (2022). Understanding the exploratory/confirmatory data analysis continuum: Moving beyond the "replication crisis." *American Psychologist, 77*(3), 453.

GAISE College Report ASA Revision Committee (2016) Guidelines for Assessment and Instruction in Statistics Education College Report. http://www.amstat.org/education/gaise.

Gelman, A., & Carlin, J. (2014). Beyond power calculations: Assessing type S (sign) and type M (magnitude) errors. *Perspectives on Psychological Science, 9*(6), 641–651.

Gelman, A., Hill, J., & Vehtari, A. (2020). *Regression and other stories*. Cambridge University Press.

Gigerenzer, G. (1987). Probabilistic thinking and the fight against subjectivity. In Kruger, L., Daston, J. L., & Heidelberger, M. (Eds.), *The probabilistic revolution, Vol. 1. Ideas in history; Vol. 2. Ideas in the sciences* (pp. 11–33). The MIT Press.

Gill, J. (1999). The insignificance of null hypothesis significance testing. *Political Research Quarterly, 52*(3), 647–674.

Google Scholar. (2024). *Google Scholar Metrics*. Google. Retrieved 08/26/24 from https://scholar.google.com/intl/en/scholar/metrics.html#metrics

Kass, R. E. (2011). Statistical inference: The big picture. *Statistical Science: A Review Journal of the Institute of Mathematical Statistics, 26*(1), 1.

Killeen, P. R. (2005). An alternative to null-hypothesis significance tests. *Psychological Science, 16*(5), 345–353.

Kish, L. (2017). Some statistical problems in research design. In Bynner, J. & Stribley, K. M. (Eds.) *Research Design* (pp. 64–78). Routledge.

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine, 15*(2), 155–163.

Kringen, J. A., Blair, J. P., & Emigh, M. (2018). *Pedagogical Issues in Criminal Justice: Often- Ignored Problems with Null Hypothesis Significance Testing ACJS Today, 43*(3), 13–22.

Lattimore, P. K., MacKenzie, D. L., Zajac, G., Dawes, D., Arsenault, E., & Tueller, S. (2016). Outcome findings from the HOPE demonstration field experiment: Is swift, certain, and fair an effective supervision strategy? *Criminology & Public Policy, 15*(4), 1103–1141.

Levchak, P. J. (2021). Stop-and-frisk in New York City: Estimating racial disparities in post-stop outcomes. *Journal of Criminal Justice, 73*, 101784.

Levine, T. R., Weber, R., Park, H. S., & Hullett, C. R. (2008). A communication researchers' guide to null hypothesis significance testing and alternatives. *Human Communication Research, 34*(2), 188–209.

Lösel, F. (2018). Evidence comes by replication, but needs differentiation: The reproducibility issue in science and its relevance for criminology. *Journal of Experimental Criminology, 14*(3), 257–278.

Lowndes, G. A. N. (1970). *The Silent Social Revolution: An Account of the Expansion of Public Education in England and Wales 1895–1965*. Oxford University Press.
Morrison, D. E., & Henkel, R. E. (Eds.) (2006). *The significance test controversy: A reader*. Transaction Publishers.
Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods, 5*(2), 241.
Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716.
Pridemore, W. A., Makel, M. C., & Plucker, J. A. (2018). Replication in criminology and the social sciences. *Annual Review of Criminology, 1*(1), 19–38.
Redelings, M. D., Sorvillo, F., Smith, L. V., & Greenland, S. (2012). Why confidence intervals should be used in reporting studies of complete populations. *The Open Public Health Journal, 5*(1), 52–54.
Rodgers, J. L. (2010). The Epistemology of Mathematical and Statistical Modeling: A Quiet Methodological Revolution. *American Psychologist, 65*(1), 1–12.
Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review, 16*, 225–237.
Rozeboom, W. W. (1960). The fallacy of the null-hypothesis significance test. *Psychological Bulletin, 57*(5), 416.
Singleton, R. A., & Straits, B. C. (2010). *Approaches to Social Research* (5th ed.). Oxford University Press.
Stunt, J., van Grootel, L., Bouter, L., Trafimow, D., Hoekstra, T., & de Boer, M. (2021). Why we habitually engage in null-hypothesis significance testing: A qualitative study. *PLoS ONE, 16*(10), e0258330.
Sweeten, G. (2020). Standard errors in quantitative criminology: Taking stock and looking forward. *Journal of Quantitative Criminology, 36*, 263–272.
Szucs, D., & Ioannidis, J. P. A. (2017). When null hypothesis significance testing is unsuitable for research: A reassessment. *Frontiers in Human Neuroscience, 11*, 390.
Tukey, J. W. (1977). *Exploratory Data Analysis*. Pearson.
Zane, S. N., Welsh, B. C., & Zimmerman, G. M. (2016). Examining the iatrogenic effects of the Cambridge-Somerville Youth Study: Existing explanations and new appraisals. *British Journal of Criminology, 56*(1), 141–160.

J. Pete Blair, Ph.D. Dr. Blair is the Director of the Advanced Law Enforcement Rapid Response Training (ALERRT) Center and a professor in the School of Criminal Justice and Criminology at Texas State University, San Marcos, Texas, USA. His work is focused on the study of active shooter events and the development of training methodologies for first responders.

Peter T. Tanksley, Ph.D. Dr. Tanksley is a research scientist at the ALERRT Center at Texas State University, San Marcos, Texas, USA. His work is focused on the health and wellness of first responders and the study of their training and behavior through experimental design.

Emily D. Spivey, Ph. D. Dr. Spivey is an assistant professor in the Criminal Justice Department at New Mexico State University. Her work is focused on reentry, criminological theory, and criminal record stigma.

M. Hunter Martaindale, Ph.D. Dr. Martaindale is the Director of Research at the ALERRT Center and a research associate professor in the School of Criminal Justice and Criminology at Texas State University, San Marcos, Texas, USA. His work integrates technologies like eye-tracking, virtual reality, and biomarkers to enhance understanding of first responders' training and behavior.