Survey Research Methods (2024)
Vol. 18, No. 3, pp. 187-210
https://doi.org/10.18148/srm/2024.v18i3.8182
European Survey Research Association

Worst Case Resistance Testing: A Nonresponse Bias Solution for Today's Survey Research Realities

Stephen L. France[1] · Frank G. Adams[1] · V. Myles Landers[1]
[1]Mississippi State University

This study proposes a method of nonresponse assessment based on meta-analytical file-drawer techniques, also known as worst-case resistance testing (WCRT), and suitable for a wide range of data collection scenarios. A general method is devised to estimate the number of significantly different nonrespondents it would take to significantly alter the results of an analysis. Estimates of nonrespondents can be plotted against effect sizes using "n-curves", with similar interpretation to p-curves or power curves. Variants of the general method are derived for tests of means and correlations. A sample using a well-established survey instrument from previous behavioral research is used to test the method. The results suggest that employing worst-case resistance testing can be used on its own or in conjunction with wave analysis to precisely flag nonresponse risks.

Keywords: participant nonresponse bias; worst-case resistance testing; hypothesis testing; wave analysis

1 Introduction

All quantitative empirical methods rely on the assumption that the sample participants represent the population of interest sufficiently to justify extrapolation of findings beyond the sample measured (Chesney & Obrecht, 2012). However, some portion of the participants solicited in almost every study do not respond, and as the proportion of those nonrespondents grows larger, the study's results suffer from potential bias (Boyd & Westfall, 1965). This participant nonresponse bias is the focus of this paper.

Participant nonresponse bias has been attributed to the variation of characteristics between respondents and nonrespondents (Deming, 1953), and this variance has the potential to confound the variance observed between the constructs measured in any given empirical test (Groves and Peytcheva, 2008) and introduce bias into statistical tests. This bias can be thought of as a type of selection bias and unlike bias for nonresponse of individual items this bias cannot be corrected without gathering data from nonrespondents (Berg, 2005).

The growing use of internet surveys for behavioral survey research has changed the nature of survey response and nonresponse. While scholars have well accepted means of assessing nonresponse bias—most notably, wave analysis—those methods were developed based upon physical mail collection of surveys. By contrast, much survey and experimental research today employs electronically curated samples that can be gathered in hours, or even minutes and that does not have well defined participant response data.

Accordingly, this study develops a set of methods independent of the survey delivery mode, allowing researchers to examine the robustness of statistical tests against participant nonresponse bias[1] by calculating the number of cases needed to reverse a statistical test over a range of different effect sizes. The resulting "n-curves" provide similar insight to related methods, such as power curves and p-curves, and provide a measure of robustness for the results of statistical tests in situations where participant nonresponse may affect the results and conclusions from such tests. Monte Carlo experiments are used to test the properties of the proposed method across multiple statistical tests. An empirical survey of customer satisfaction is then used to show how these methods can be used on their own or combined with wave analysis to flag statistical tests where nonresponse bias may be problematic. The data files and code used in the creation of this paper have been made available (France et al., 2024a, b).

Corresponding author: Stephen L. France, Mississippi State University, Mississippi State, MS 39762, USA (Email: sfrance@business.msstate.edu)

[1] For the sake of parsimony, as this paper focuses on participant nonresponse issues, subsequent mentions of nonresponse refer to participant nonresponse rather than item nonresponse (e.g., Skafida et al., 2022).

Table 1

Summary of Nonresponse Bias Assessments and Remedies

Technique	*Description*	*References*
Comparison of Sample and Population	Compare demographics of known population characteristics to collected sample characteristics	Armstrong and Overton (1977); Groves (2006); Beebe et al. (2011)
Wave Analysis	Compare early respondent answers to later respondent answers	Armstrong and Overton (1977)
Follow-up Analysis	Obtain responses from subjects who did not respond to the original data collection in order to test for differences	Aiken (1981), Sosdian and Sharp (1980)
Bayesian Analysis	Utilize Bayes rule to estimate nonresponse data, assuming independence of attributes and known characteristics across respondents and nonrespondents	Daniel and Schott (1982)
Passive and Active Nonresponse Analysis	Attempt to assess why active nonrespondents declined to participate through focus groups, interviews, and surveys about the original data collection. Resend the survey to address passive nonrespondents	Rogelberg and Stanton (2007); Rogelberg et al. (2003); Roth (1994)
Interest-level Analysis	Include questions about subject interest in the survey topic among the measured items and statistically control for interest when analyzing responses	Rogelberg and Stanton (2007); Rogelberg et al. (2000)
Benchmarking	Compare sample demographics with those of other studies of similar phenomena to see if there are inconsistencies of means or standard deviations	Rogelberg and Stanton (2007)
Replication	Conduct multiple surveys using different samples to assess whether findings remain consistent	Rogelberg and Stanton (2007)
Weighting	Add additional covariates and use correlations of these covariates with nonresponse to weight responses	Wetzel and Hünteler (2022)
Selection Bias Function	Creation of a selection bias function based on expert judgments of feasible variable ranges to account for missing data in parameter estimates	Rotnitzky et al. (1998); Scharfstein and Irizarry (2003)
Propensity Score Analysis Weighting	Utilizing propensity score analysis to weight observations using covariate information	Lee (2006); Schonlau et al. (2009)
Manski Bounds	Calculate minimum and maximum values of the dependent variable given a feasible range for missing data and utilize these estimates to create bounds on the dependent variables	Horowitz and Manski (1998); Manski (2016)

2 Background

In recent years, dedicated efforts have examined practices such as “p-hacking” to recall the academy to replicable research methods (Simmons et. al, 2011). Similarly, recent failures to replicate psychological research have been a cause for concern (Stanley et al., 2018), leading to a call for more transparency in research (Inman et al., 2018). This includes the reporting of data collection techniques, statistical power, effect sizes, and potential biases that might influence results.

Arising from a combination of sampling error and coverage error, nonresponse error results in a sufficient difference between the data sought by a researcher and the data actually obtained to compromise a study’s validity (Collier & Bienstock, 2007). Conceptually, nonresponse error holds that the potential responses of subjects who do not answer a solicitation to participate in a given research study might be different enough from the responses recorded to alter the findings of the study and higher nonresponse rates can negatively affect the representativeness of a sample (Cook et al., 2000). When records of response data are available, it is quite easy to assess participant nonresponse, but “there is no magical response rate below which an observed mean, standard deviation, or correlation becomes automatically invalid” (Newman 2009, pp. 7). Still, the larger the percentage of the solicited sample measured, the lower the error resulting from nonresponse bias tends to be (Olson, 2006).

Scholars have developed several procedures to adjust survey results to account for survey nonresponse bias, as detailed in Table 1 (following Halbesleben & Whitman, 2013), but they all inherently rest on a key assumption: “... respondents and nonrespondents within a weighting class have the same values on key variables ...” (Groves 2006, pp. 653). Accordingly, attempts to assess nonresponse bias rely on an assumption that differences causing some solicited subjects to forego answering a survey are related to how a nonrespondent might react to a study’s constructs of interest. Based on this assumption, for decades, the most widely used method of assessing nonresponse bias has been Armstrong and Overton’s wave analysis technique (1977).

2.1 Covariate Methods

Several of the listed methods in Table 1 utilize covariate information in order to account for differences between the sample and the solicited respondents. Methods include weighting using correlations with covariates (Wetzel & Hünteler, 2022), propensity score methods (Lee, 2006; Schonlau et al., 2009), and methods that utilize feasible ranges of covariate information to create selection bias functions for parameter estimation (Rotnitzky et al., 1998, Scharfstein & Irizarry, 2003) and for creating solution bounds (Horowitz & Manski, 1998; Manski, 2016). These methods can be particularly useful in longitudinal data collection, where subject covariates are known, but subjects may miss measurement occasions or drop out of an experiment.

2.2 Wave Analysis

Simply put, wave analysis compares relationships between variables observed among early respondents to a measurement instrument with those observed among later respondents (Armstrong & Overton, 1977). “The basic assumption ... is that subjects who respond less readily are more like those who do not respond at all than those who do respond readily (i.e., those who respond sooner and those who need less prodding to answer)” (Kanuk & Berenson, 1975, pg. 449). The method poses that a lack of significant difference between early and late respondents to a research solicitation implies that potential subjects that did not respond do not represent observations that might alter an analysis’s results.

For all its long-proven utility (over 21,000 citations at this writing), wave analysis was explicitly built around mail surveys, which generally require considerable periods of time to collect (Kanuk & Berenson, 1975) and where information on early and late response waves can easily be found. Studies employing postal mail and citing wave analysis have included follow up prompts of up to four weeks (Diamantopoulos & Winklhofer, 2001; Mohr & Spekman, 1994; Sirdeshmukh et al. 2002). Even surveys distributed over email have noted significant time spent waiting for responses from subjects (Pavlou, 2003). As of 2020, the vast majority of surveys were completed via the internet (Daikeler et al., 2020). The “internet age” of surveys has seen a growth of third-party survey platforms, such as Qualtrics and Prolific, who recruit participants well in advance of any study, and pay participants fees to complete studies. The resulting participants are more likely to reply quickly because they have pre-agreed to participate in studies (Qualtrics, 2020). On some research platforms, such as Prolific (Peer et al., 2017) and the Amazon Mechanical Turk (Chandler et al., 2019), potential respondents, when logging on, will pick from a list of potential surveys or work to complete. In this situation, unless user click/screen viewing behavior is analyzed, it is difficult to identify and quantify nonresponses (e.g., Boas et al., 2020; Paolacci et al., 2010) and the early and late response waves required by wave analysis.

2.3 Resampling

A typical means of addressing potential nonresponse bias is to simply resolicit sample nonrespondents (Aiken, 1981; Hartman et al., 1986; MacDonald et al., 2009), often employing shorter surveys that assess only the items whose constructs are of critical importance to the observed findings, to look for differences from the findings of the original survey (Lambert & Harrington, 1990). However, the absence of a specific response rate below which nonresponse bias is considered problematic (Newman, 2009) implies that supplemental sampling—whether among the originally solicited group, or from a different group of potential respondents—may not necessarily address nonresponse bias of a sample relative to the population of interest. A different potential solution may lie in using meta-analysis techniques to address a bias issue known as the file drawer problem.

2.4 Meta-Analysis and the File Drawer Problem

The file drawer problem is a term used in meta-analytic literature to describe a conceptual, but quantifiable sampling bias. Because meta-analyses examine the standardized results of extant literature, they are presumed to be biased by the tendency of statistically significant findings to achieve academic publication, and the corollary tendency of nonsignificant results of similar phenomena never entering the scholarly body of knowledge (Rosenberg, 2005). The worst assumptions hold that 95% of contrary findings do not survive the academic publication process, and that the body of knowledge is, therefore, a victim of Type 1 error (Rosenthal, 1979).

Rosenthal (1979) proposed a solution to the file drawer problem, sometimes known as worst-case resistance testing, or fail-safe number calculation (Rosenberg, 2005). The technique calculates the number of studies required to significantly alter an observed mean of effect sizes, assuming the hypothetical unobserved studies have a collective mean significantly different than that of the observed effect sizes. As this calculated number of studies increases, the likelihood of a file drawer bias decreases. In other words, the larger the effect size observed in tests of a given sample, and/or the less stringent the standard of testing significance, the more hypothetical contradictory cases it would take to cast doubt on the observed findings. The technique has been used in varying meta-analytic studies including (but by no means limited to) electronic word of mouth (Babić et al., 2016), interstitial space impacts on consumer appeal (Sevilla & Townsend, 2016), and consumer responses to humanoid robots (Mende et al., 2019).

The nonresponse bias problem is very similar to the file drawer problem in that both seek to assess a difficult-to-quantify bias of findings stemming from uncollected data presumed to contradict results based on observed data. It stands to reason that the file drawer or worst-case resistance testing (WCRT) solution should also be efficacious in assessing nonresponse bias.

3 Methodology

To illustrate how file drawer concepts can be applied to the nonresponse bias problem, the problem is given in general in terms of the basic NHST (null hypothesis significance test) paradigm. Though this paradigm has been much criticized (e.g., Gill, 1999; Hubbard & Armstrong, 2006; Hunter, 1997; Schneider, 2015) it is still by far the predominantly used framework for building theory in empirical management and social science research.

Furthermore, most alternative approaches proposed to replace NHST also have criticisms. For example, the use of confidence intervals for inference leads to the same "inverse inference" that is criticized in NHST testing, and Bayesian analysis requires specification of prior distributions, which can be conceptually difficult (e.g., Trafimow, 2017). While at least one journal has banned significance testing (Woolston, 2015), most journals and scientific associations in the behavior sciences and business disciplines have focused on best practice to improve the use of NHST results and to put these results into context.

Scholars have advanced several recommendations to improve the implementation of NHST methods. These include putting p values into context and avoiding erroneous overly strong conclusions from p values (Wasserstein & Lazar, 2016), focusing on the magnitude and size of any statistical effect and incorporating information from prior beliefs using Bayes factors (Harvey, 2017; Valentine et al., 2019), reporting of descriptive statistics and reporting guidelines for major statistical tests (JCR, 2021), including detailed graphs and discussions of effects and utilizing robust error statistics (Schwab et al., 2011), and calculating power values for each statistical test and ensuring that the Type II error rate (β) is less than 0.05 when making conclusions on a lack of "effect" relative to a null hypothesis (Baroudi & Orlikowski, 1989; Cashen & Geiger, 2004).

A theme in most of the rules and suggestions described above is the "triangulation" of NHST results with other metrics to build evidence for hypothesis test conclusions. As such, the methodology described in this paper fits in with this theme. The aim is to provide a set of measures of robustness of statistical results to problems caused by nonresponse bias. However, the methods described can be used beyond the realm of nonresponse bias to examine robustness to other sources of error, such as the experimental design.

In this study, the WCRT methodology is described using a generic NHST hypothesis testing procedure. Examples are included for problems with simple hypothesis testing of means and of correlations, where equations are given for "finding the number of additional studies" required to negate a conclusion and then models are developed to solve these equations.

3.1 The General Model

The general problem is outlined as follows: Consider a situation with a NHST performed on data collected from a survey. The purpose of the test is to find sufficient evidence to reject a null hypothesis (H_0) in favor of an alternative hypothesis (H_A). There is some critical value at which enough evidence is gathered so that the researcher flips from failing to reject the null hypothesis to rejecting the null hypothesis. If the researcher finds enough evidence to reject H_0, but H_0 is in fact true, then the researcher is considered to have committed a Type I error with a probability denoted as α. The value of α is usually defined in terms of extreme results in the distribution of expected sample values in the H_0 distribution, which can be denoted as $\alpha = P\left(R \mid H_0\right)$, where R is rejection of H_0. Given the distribution of H_0, H_0 is rejected if there is enough evidence, operationalized by the sample statistic being far enough away from a "null effect" in a sampling distribution.

A researcher will often make the "opposite assertion", that given insufficient evidence to reject H_0, one can conclude that H_0 is in fact true. However, there is a danger with this assertion in that researchers may assume a trivial effect without understanding the implications of the power of the statistical test (Baroudi & Orlikowski, 1989; Cashen & Geiger, 2004; Sawyer & Ball, 1981). If the researcher fails to reject H_0 and in fact H_A is true, then the researcher has made a Type II error, i.e., $\beta = P\left(R^c \mid H_A\right)$, where the power of the test is $1–\beta$. An issue here is that H_A can take multiple values and that the power varies with the "effect size" difference between H_0 and the H_A used to calculate power. Solutions to this issue include calculating power using a reasonable effect size based on prior studies, standard small, medium, and large effect sizes (Cohen, 1992), and g raphing power values across a range of effect sizes, a “so called” power curve (Faul et al., 2007).

In the context of nonresponse bias and WCRT, the focus is to find the number of nonrespondents who can reverse a statistical conclusion and use this as a measure of robustness of the solution. But how is the effect size for these studies chosen? Is it a “zero effect”, the opposite effect, or a smaller effect in the same direction? The methodology outlined in this paper mirrors the work described above in choosing effect sizes for power analyses. The number of nonrespondents needed to reverse a statistical test can be calculated for a range of feasible effect sizes, which can be estimated from wave analysis or by examining effect sizes for similar studies. These values can be plotted, creating an “n curve”, which is similar to curves used for determining quality bounds for confidence intervals (e.g., Trafimow, 2018) or *p*-curves used to map sample sizes for different *p* values at different power levels (Simonsohn et al., 2014).

At the core of the analyses in this paper is the idea of a standardized effect size (Cohen, 1988). An effect size can be thought of as a quantitative measure of the phenomenon being studied (Kelley & Preacher, 2012). For example, for a single sample t test, the effect size *d*, is given in (1).

$$d = \frac{\overline{x} - \mu_0}{s} = \frac{(\overline{x} - \mu_0)/\sqrt{n}}{s/\sqrt{n}} = \frac{t}{\sqrt{n}}, \qquad (1)$$

where $\overline{x}$ is the sample mean, s is the sample standard deviation, μ_0 is the hypothesized population mean, and n is the sample size. This invariance towards n is particularly useful for large sample size experiments, as effect sizes can put into context results that are statistically significant with only a small effect size, but a very large sample size (e.g., Coe, 2002). Different statistical tests have different effect size calculations. For example, effect sizes for the comparison of two group means, such as Cohen’s d and Hedge’s g, have a similar format to the effect size given in (1), while for Pearson’s correlation, the sample regression coefficient *r* is often used as a measure of effect size (Hemphill, 2003).

In the context of a WCRT analysis of a NHST test, we define a general effect size φ, which can be substituted by the appropriate metric for a specific test (e.g., d for a sample mean test). Consider the following situations:

1. With a sample size of size n_1 and effect size φ_1 there is enough evidence to reject H_0. We wish to find n_2, where this is the number of items or nonrespondents with effect size φ_2 required to negate the result, so that H_0 is no longer rejected.
2. With a sample of size of n_1 and effect size φ_1, there is not enough evidence to reject H_0. We wish to find n_2, where this is the number of items or nonrespondents with effect size φ_2 required to negate the result, so that H_0 is now rejected.

A set of candidate effect sizes needs to be defined for φ_2. This is key to the methods described in this paper and the appropriate range can be informed by previous research, the results from a wave analysis of the data, and the effect size φ_1 (for example, if there is a significant effect, an effect size greater or equal to φ_1 and in the same direction is not going to negate the hypothesis test). For each $\varphi_{2,}$ the procedure will give the n_2 value needed to reverse the result of the statistical test.

3.2 Inference for Single Sample t test

Consider a single sample t test of a population mean being equal to hypothesized mean μ_0. The notation is as per (1) and the null hypothesis is $H_o : \mu = \mu_0$. The methodology outlined in this section covers both two tailed tests where the alternative hypothesis is defined as $H_a : \mu \neq \mu_0$ and one-tailed tests where the alternative hypothesis can be defined as $H_a : \mu > \mu_0$ or $H_a : \mu < \mu_0$. The test statistic derived from the sampling distribution is defined as (2) by rearranging (1).

$$t = \frac{(\overline{x}_1 - \mu_0)}{\frac{s_1}{\sqrt{n_1}}} = d_1\sqrt{n_1} \qquad (2)$$

Here, the subscript 1 indicates that the sample values are based on the responses, while the subscript 2 will be used for the sample values for hypothesized nonresponses. The t distribution has n-1 degrees of freedom and varies with n. Let t^* be the critical boundary between rejecting and failing to reject H_0. Dependent on n and the strictness of the test (using the Type I error α), H_0 is rejected if $|t| > t^* = t_{\alpha/2}$ for a two-tailed test and $t > t^* = t_\alpha$(or $t < t^* = -t_\alpha$) for a one-tailed test.

Here we consider four different scenarios[2].

1. For a one-tailed upper test or a two-tailed test with $t > 0$, H_0 is rejected as $t > t^*$. We wish to find, for some *nonrespondents* with effect size d_2, the n_2 for required to reverse this conclusion, so that $t \leq t^*$.
2. For a one-tailed upper test or a two-tailed test with $t > 0$, H_0 is not rejected as $t \leq t^*$. We wish to find, for some *nonrespondents* with effect size d_2, the n_2 required to reverse this conclusion, so that $t > t^*$.

[2] The only scenario not covered by the above is the “so-called” type III error scenario (Leventhal & Huynh, 1996), where the sample mean is in the opposite direction to the population mean.

Table 2

Scenarios for Single Sample t test

Scenario	*Test Direction*	*Test Result*	*Find min{n_2} to make*	ε	*Nonresponse d_2 Bounded*
1	Upper	Significant	Non-significant	$\varepsilon \geq 0$	Upper
2	Upper	Non-significant	Significant	$\varepsilon < 0$	Lower
3	Lower	Significant	Non-significant	$\varepsilon \leq 0$	Lower
4	Lower	Non-significant	Significant	$\varepsilon > 0$	Upper

3. For a one-tailed lower test or a two-tailed test with $t < 0$, H_0 is rejected as $t < t^*$. We wish to find, for some *nonrespondents* with effect size d_2, the n_2 required to reverse this conclusion, so that $t \geq t^*$.
4. For a one-tailed lower test or a two-tailed test with $t < 0$, H_0 is not rejected as $t \geq t^*$. We wish to find, for some **nonrespondents** with effect size d_2, the n_2 required to reverse this conclusion, so that $t < t^*$.

A simplifying assumption is to assume that $s_1 = s_2$, i.e., the nonresponse and response data have the same standard deviations. However, if the nonresponse data have different characteristics than the original data then this assumption will not hold. A solution is to set some range for s_2, so that $(1-\theta)\, s_1 \leq s_2 \leq (1+\theta)\, s_1$, where $0 \leq \theta \leq 1$ and θ is set based on some prior inferences regarding the data. Given s_2, the effect size for the nonresponse data is given in (3).

$$d_2 = \frac{\overline{x}_2 - \mu_0}{s_2} \tag{3}$$

The sample mean for the nonresponse data is found by rearranging (3) to give (4).

$$\overline{x}_2 = d_2 s_2 + \mu_0 \tag{4}$$

This value can be used to find the sample mean for the combined response and nonresponse samples.

$$\overline{x}_c = \frac{n_1\overline{x}_1 + n_2\overline{x}_2}{n_1 + n_2} \tag{5}$$

The pooled standard deviation can be calculated using the meta-analysis formulation given in Higgins et al. (2019).

$$s_c = \sqrt{\frac{(n_1-1)\, s_1^2 + (n_2-1)\, s_2^2 + \frac{n_1 n_2}{n_1+n_2}\left(\overline{x}_1^2 + \overline{x}_2^2 - 2\overline{x}_1\overline{x}_2\right)}{n_1 + n_2 - 1}} \tag{6}$$

Consider the overall t test with the combined data for scenario 1. We wish to find the lowest n_2 for which $t \leq t^*$, and define some small quantity ε, such that $t + \varepsilon = t^*$, with $\varepsilon \geq 0$, so that $t = t^* - \varepsilon$.

$$t^* - \varepsilon = \frac{\overline{x}_c - \mu_0}{\left(\frac{s_c}{\sqrt{n_1+n_2}}\right)} \tag{7}$$

Utilize to expand out $\overline{x}_c$ in terms of n_1 and n_2.

$$t^* - \varepsilon = \frac{\frac{n_1\overline{x}_1 + n_2\overline{x}_2}{n_1+n_2} - \mu_0}{\left(\frac{s_c}{\sqrt{n_1+n_2}}\right)} \tag{8}$$

Multiply both the numerator and denominator by $n_1 + n_2$ and then rearrange the numerator terms.

$$t^* - \varepsilon = \frac{n_1\left(\overline{x}_1 - \mu_0\right) + n_2\left(\overline{x}_2 - \mu_0\right)}{s_c\sqrt{n_1 + n_2}} \tag{9}$$

The task is to find the smallest integer value of n_2 for which $\varepsilon \geq 0$. By making minor alterations, (9) can be used for scenarios (2)–(4). The four scenarios are summarized in Table 2.

For each scenario, Table 2 gives the test direction (upper or lower), the result of the test on the response data, the opposite result, the range of ε for which the minimum integer n_2 is being found, and how the nonresponse effect size is bounded[3]. Given that the t-value in (9) is dependent on n_2, giving a cross-dependency, n_2 cannot be calculated directly. A fixed-point optimization procedure for finding n_2 is given in Appendix A. An almost identical procedure can be used for the single sample z test.

A similar process can be followed for two-sample independent sample tests. Sample sizes can be calculated for Student's t test (for equal variances), Welch's t test (for unequal variances), and the two sample z test. A two-group measure, such as Hedge's g or Cohen's d can be used to calculate effect sizes (Rosenthal & Rubin, 1982). If sample sizes are uneven, some constraints need to be placed on rel-

[3] The exact bounds are not given here as there is a nonlinear dependence between sample size and effect size. For scenarios 1 and 4, there is an upper bound on effect size at which n_2 goes to infinity. For scenarios 2 and 3, there is a lower bound on effect size at which n_2 goes to infinity.

ative group sizes. Derivations for the two-sample tests are included in Appendix B.

3.3 Inference for Correlation Test

Consider a situation where a correlation is being tested for significance. The null hypothesis is $H_o : \rho = 0$, where ρ is the population correlation. Standard alternate hypothesis are $H_a : \rho > 0$ (or $\rho < 0$) for a one-tailed test and $H_a : \rho \neq 0$ for a two tailed test. A population hypothesis is tested with a Pearson sample correlation coefficient r. The correlation r is essentially an effect size (Cohen, 1988), with small ($0.1 \leq r < 0.3$), medium ($0.3 \leq r < 0.5$), and large ($r \geq 0.5$) effect sizes defined.

There are several different tests for the significance of correlations. The one most commonly used in meta-analysis involves transforming the correlation $r \in [-1,1]$ into a z score using the inverse hyperbolic tangent transformation (Cox, 2008) and is given in (10).

$$zr_1 = \tanh^{-1}(r_1) = \frac{1}{2}\ln\left(\frac{(1+r_1)}{(1-r_1)}\right), \tag{10}$$

where r_1 is the correlation coefficient for the response data. Now, this value is still essentially an effect size and does not depend on the sample size n. A standard error of $\sqrt{1/(n-3)}$ is defined by Fisher (1921), which can be used to give the z statistic in (11).

$$z_1 = \frac{zr_1}{\mathrm{SE}(zr_1)} = \frac{zr_1}{\sqrt{1/(n-3)}} \tag{11}$$

Given that the z test is a simple two-way directional test, the four scenarios for finding the n_2 values needed to change a hypothesis test result are similar to the scenarios outlined for the one sample t test in Table 2. The only changes are that “z” replaces “t” for the test statistic and critical values, and that the effect size defined for the nonresponse data is a correlation coefficient r_2, which can be transformed into a z score zr_2 using the transformation given in (10). The z-scores for the response and hypothesized nonresponse data can be combined (Field, 2001; Hedges & Vevea, 1998; Higgins et al., 2019) using (12).

$$zr_c = \frac{(n_1-3)\,zr_1 + (n_2-3)\,zr_2}{n_1+n_2-6}, \tag{12}$$

which has the standard error given in (13).

$$\mathrm{SE}(zr_c) = \sqrt{\frac{1}{n_1+n_2-6}} \tag{13}$$

For scenario 1, we wish to find the lowest n for which $z \leq z^*$, where z^* is the boundary value for significance and define some small quantity ε, such that $z + \varepsilon = z^*$, with $\varepsilon \geq 0$, so that $z = z^* - \varepsilon$.

Table 3

Monte Carlo Experiment Factors

Factor and Factor Levels	*Description*
Method zSingleSample, tSingleSample, zTwoSample, tStudentTwoSample, tWelchTwoSample, zCorrelation	The method tested. Each of the methods described in the previous section was tested
nSample 25. 100, ..., 500	The sample size for the data, from 25 to 500 with increments of 25
Alpha 0.1, 0.05, 0.01	The alpha (α) Type I error of the statistical test
Distribution Normal, Uniform, Poisson, NExponential	The error distribution the data are sampled from. This tests the robustness of results to data that do not necessarily conform to the assumptions of statistical tests. All distributions are scaled to give a mean of 0 and a standard deviation of 1
EffectSize Low(-ve), Medium(-ve), High(-ve), Low(+ve), Medium(+ve), High(+ve)	The average effect size of the generated data. A higher effect size indicates a stronger likelihood of significance. The effect sizes are the standard effect sizes defined by Cohen (1988, 1992), i.e., for mean tests $d = (0.2, 0.5, 0.8)$, and for correlations $r = (0.1, 0.3, 0.5)$
EffectSize2 Low, Medium, High	The WCRT file draw effect size in the opposite direction to the sample data effect size

$$z^* - \varepsilon = \frac{zr_c}{\mathrm{SE}\left(zr_c\right)} = \frac{zr_c}{\sqrt{\frac{1}{n_1+n_2-6}}} \tag{14}$$

This equation can be rearranged to give n_2.

$$n_2 = \left(\frac{z^* - \varepsilon}{zr_c}\right)^2 - n_1 + 6 \tag{15}$$

Now, n_2 can be found in a similar manner to the single sample hypothesis test. The other three scenarios can be taken from Table 2 (with r replacing d in the final column). A local search optimization procedure for finding n_2 is given in Appendix A.

4 Monte Carlo Simulations

To help validate the methods described in the previous section, several Monte Carlo simulation experiments were run. The methods outlined in this article are not "supervised" in the traditional sense, in that they give an analytical measure of resiliency rather than an ex-ante prediction of a dependent variable. However, the methods can be evaluated in a similar manner to evaluations of unsupervised analysis, for example, cluster analysis. Here, the method assumptions and the results generated by the method are tested on error-perturbed input data, to see how the resilient the model is to error. This approach has been widely used in dimensionality reduction (e.g., Akkucuk & Carroll, 2006; Chen & Buja, 2009) and cluster analysis (e.g., Banfield & Raftery, 1993; Brusco et al., 2017; Milligan, 1981). For example, in cluster analysis, one can generate items for clusters with Gaussian error from the cluster centroids and then see how well items are assigned to the correct clusters. We followed this approach and generated a range of data distributions for each of the tested WCRT methods, varying the factors of the sample size, the error distribution, the test coefficient, the sample data effect size, and the opposing WCRT (nonrespondent) effect size. The factors and factor levels in the experimental design are summarized in Table 3.

Regardless of the statistical test, the general properties of the WCRT methodology outlined in the previous section should give estimates with the following causal relations for significant statistical tests (in each case holding all other variables constant).

- Property 1: As the sample size n increases, the WRCT n should increase (as more evidence is required to reverse the test result).
- Property 2: As the value of α increases, the WRCT n to reverse the result should increase (as hypothesis test boundary is closer to the center of the H_0 distribution).
- Property 3: As the overall sample data effect size for the sample increases, the WRCT n to reverse the result should increase (as the sample effect is stronger).
- Property 4: As the file drawer WRCT (opposing) effect size increases, the n to reverse the result should decrease (as the items in the opposing sample are more strongly opposing the test result).

In a full factorial design, for each combination of factor levels, 10 random samples were taken, giving a total of 259,200 experimental runs. For each solution, an effect size was taken and then random error was added for each of the items in the generated data. The random error was generated from the stated distribution (normal, uniform, Poisson, or negative exponential) and scaled to give a mean of 0 and standard deviation of 1 (as the error is relative to a standardized effect size).

As an initial test of experimental procedure and as a manipulation check, the proportion of significant tests was summarized for each combination of test and effect size (small, medium, and large). The resulting bar graph is given in Fig. 1. As the proportion of significant results increases

Table 4

OLS Regression on Monte Carlo Simulation

			95% C.I.	
IV	*Coef.*	*Std. Err.*	*Upper*	*Lower*
Test				
Singlet	2.9*	1.28	0.4	5.5
Singlez	2.9*	1.28	0.4	5.4
TwoStudentt	–31.5***	1.30	–34.1	–29.0
TwoWelcht	–31.4***	1.30	–34.0	–28.9
Twoz	–30.7***	1.30	–33.3	–28.2
Sample effect				
Medium	208.2***	0.95	206.3	210.0
Large	408.8***	0.95	407.0	410.7
n (sample)	1.2***	0.00	1.2	1.2
WCRT Effect				
Medium	–270.2***	0.89	–271.9	–268.4
Large	–343.5***	0.89	–345.3	–341.8
Alpha				
0.05	51.7***	0.91	50.0	53.5
0.1	80.8***	0.90	79.1	82.6
Intercept	–123.4***	1.64	–126.7	–120.2
Observations		225,635		
R^2		0.70		

*$p<0.05$, **$p<0.01$, ***$p<0.001$

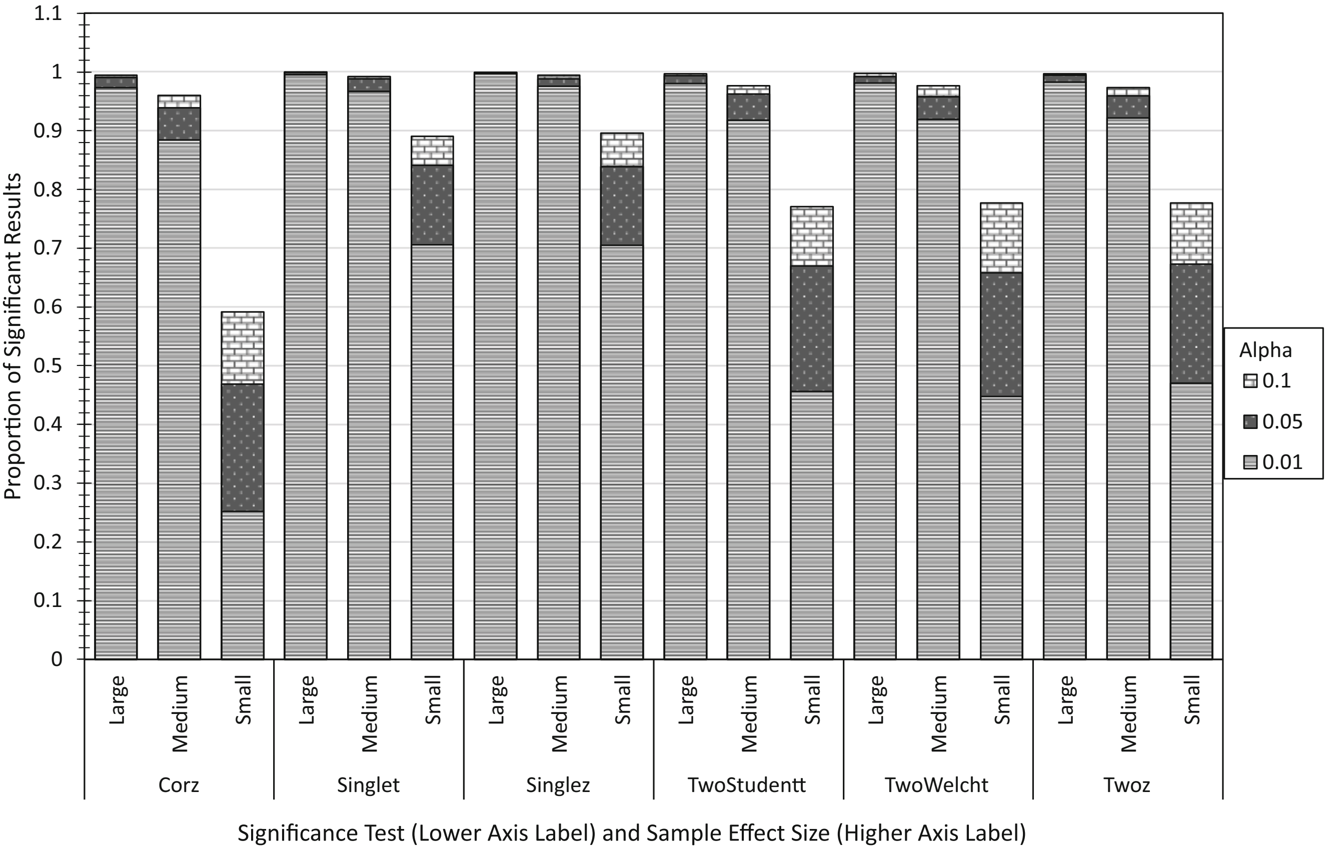

Fig. 1

The Proportion of Significant Results by Test, Effect Size, and Alpha (α)

as the hypothesis tests become less strict with higher α, results are overlaid for the different values of α, with the smallest values of α plotted at the front. As expected, for every test, the proportion of significant results is monotone increasing with the sample data effect size.

Further analyzing the significant test results (n = 225,635), summary graphs were produced to illustrate the properties outlined in points 1–4. Fig. 2 gives the WRCT n required to reverse significance given across all combinations of sample data effect size and opposing WRCT effect size. Each bar is for a WRCT file drawer effect size (higher axis description) and three values of α. As per Fig. 1, lower values of α are plotted at the front. The bars are grouped by the sample data effect size (lower axis description). Once can see that as alpha (α) increases, the WRCT n increases, giving evidence for Property 2. As the sample effect size increases, the WRCT n increases, giving evidence for Property 3, and as the WCRT file drawer effect opposing effect size increases, the WRCT n decreases, giving evidence for Property 4.

Fig. 3 gives the WRCT n required to reverse significance given across all values of alpha (α) and all combinations of data sample size. As with Fig. 2, as alpha (α) increases, the WRCT n increases, giving evidence for Property 3. There is a positive relationship between the data sample size n and the WRCT n, giving evidence for Property 1. From the previous analytic work, the relationship should be linear, as the WRCT n is linearly related to the sample size n. There is a small deviance from linearity due to the error added to the data, but the overall relationship still strongly holds.

To further understand the effect of the experimental factors, an OLS regression was run with the WRCT n as the dependent variable and the experimental factors as the independent variables. As the sample data effect could be either positive or negative, this variable was split into the direction (positive or negative) and the size of the effect. A regression with main effects and a regression that also included two factor interactions were run. The addition of two-factor interactions only improved the model fit slightly, so for the sake of parsimony the simpler main effect model was reported. All variables apart from the direction and error distribution were significant. These variables were removed from the model. For the direction, this behavior is expected as a negative data sample effect and a positive

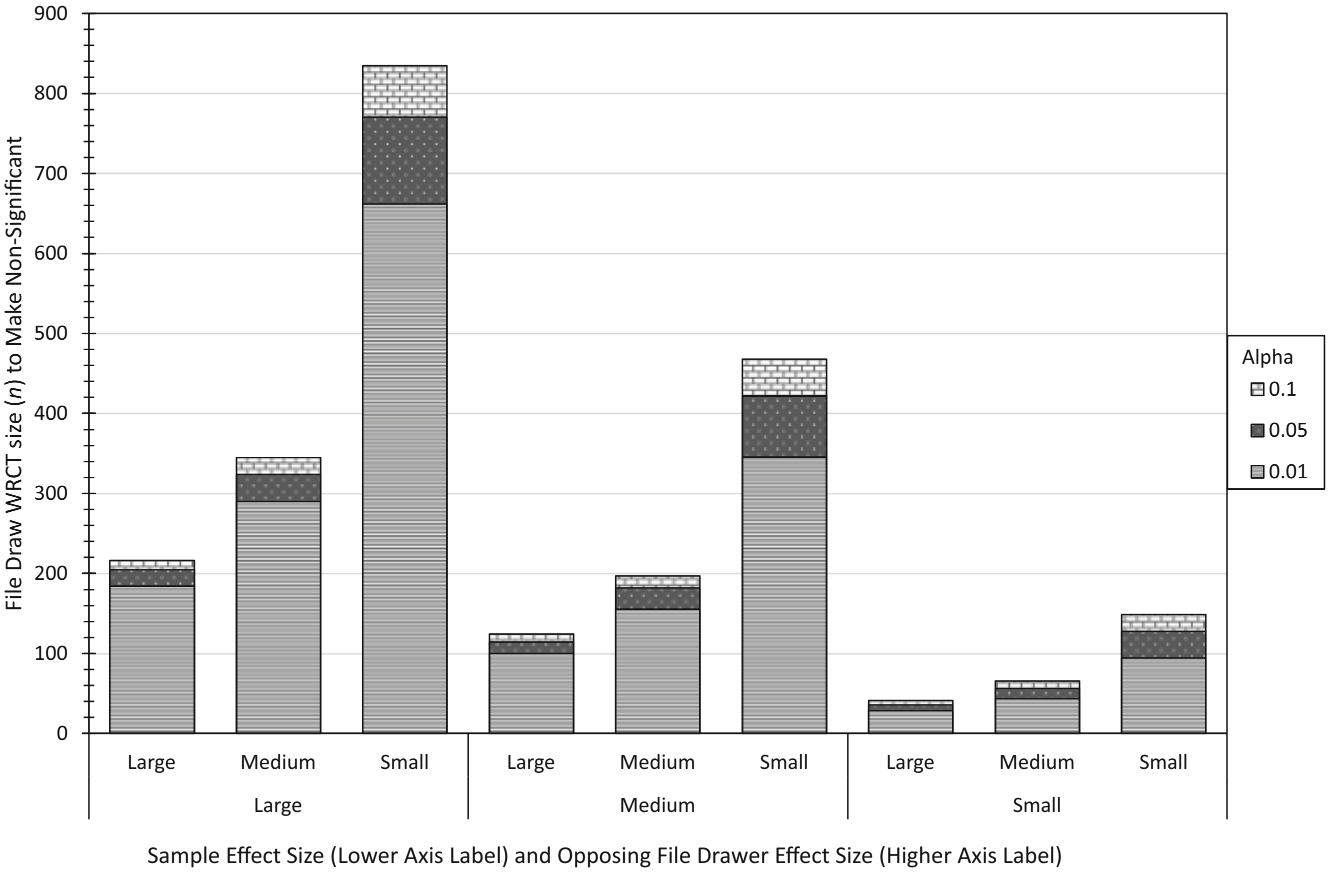

Fig. 2

The File Drawer Number of Studies to Negate Significance Across Different Effect Sizes and Alpha (α) Values

WCRT effect size should give the same WCRT n as a positive data sample effect and a negative WCRT effect size with the same magnitudes. The different error distributions were tested to show the robustness of the WCRT procedure with different types of error. However, these do not strongly affect the performance. It may be that for larger sample sizes the central limit theorem makes the actual error distribution unimportant. The results are reported in Table 4. The reference levels are "Small" for the sample data and WRCT effect sizes, alpha ($\alpha = 0.01$) for the significance level, and the Correlation z test for the test. The overall R^2 is 0.6989, indicating a strong fit, through with some variability due to the error added to the sample datasets.

In line with the model free evidence in the graphs, the WRCT n increases as the sample size increases, as the sample data effect size increases, and as the alpha increases. This gives additional evidence for Properties 1 to 4. The statistical test factor is significant, but this only gives evidence that the different statistical tests require different WRCT n values to reverse significance relative to the other experimental factors, which is to be expected. Overall, the tests show that the WRCT procedure works as expected and is robust to error variance in the sample data.

4.1 Simulation Experiment for Regression Data

This section demonstrates the use of the framework for regression. A real world dataset was taken from Hernán & Robins (2020), which contains cleaned data for the National Health and Nutrition Examination Follow-up Study (NHEFS). The data can be utilized for regression analysis, where the purpose is to explain and predict a subject's cholesterol level from a range of demographic and health indicators, including height, weight, blood pressure, gender, age, income, smoking habits, alcohol consumption, and exercise levels. The utilized dataset in total had 12 predictors and 1461 complete rows of data. Here, the fact that OLS

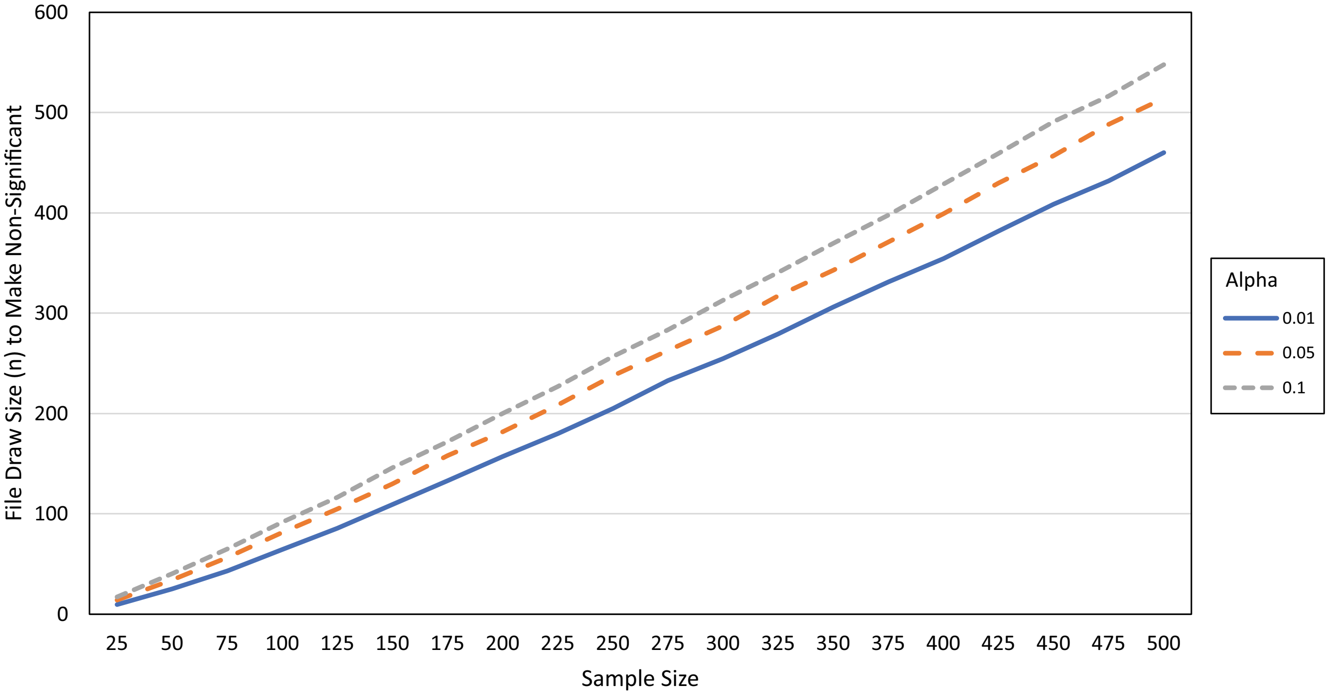

Fig. 3

The File Drawer Number of Studies to Negate Significance Across Different Sample Sizes and Alpha (α) Values

regression generates a correlation variable is exploited to give a measure of robustness for OLS regression[4].

It can be assumed that as more predictors are added to an OLS regression model the R^2 value and thus the correlation (r) will increase. As with the previous experiment, larger values of the (opposite) WRCT effect size will give smaller WRCT n robustness values and larger values of the Type I error (α) will give larger WRCT n robustness values. To test these properties, the model was built with each possible combination of predictors. To simulate potential sources of error variance, each dataset was sampled with replacement in a similar manner to how bootstrapping is used to measure error variance on survey data (e.g., Sitter, 1992). In total, given the twelve regression predictors, three potential values of the WRCT effect size (all $r = \sqrt{R^2}$ are positive measures of effect size, so only negative small, medium, and large opposing WRCT effects were used), and three values of α, there were $2^{12} \times 3 \times 3 = 36{,}864$ experimental conditions. Each experimental condition was run with five replications. The resulting correlation and the opposing WRCT effect size were used to calculate the WCRT n measure of robustness.

All experimental runs apart from runs with models with no predictors and models with only income as a predictor were significant. The significant results aggregated across the number of predictors and the opposing WCRT effect size (r_2 = –0.1 (small), –0.3 (medium), –0.5 (large)) are given in Fig. 4. Here, as expected, the WRCT n measure of robustness increases on aggregate with the number of predictors and decreases with the size of the opposing WCRT effect size (with larger effect sizes plotted at the front).

To further test the effect of the predictors, an OLS regression was run with the WCRT (n) as the dependent variable and the predictor presence (0 or 1) for the regression predictors, Type I error (α), and WCRT effect size as independent variables. As expected (Property 4), and in line with Fig. 4, the WRCT (n) is monotone decreasing with the WRCT effect size. In addition, as the value of α increases, making the tests less strict, the value of WRCT (n) required to reverse the tests also increases (Property 2). Every single one of the presence variables is positive, which is in line with expectations, as each additional variable added has a positive effect on the model fit (e.g., one can always set a variable coefficient to 0 and get an identical fit to the model without the variable). Overall, the R^2 of the model is 0.8665, indi-

[4] The WRCT (n) measure can be utilized as a measure of robustness. However, the Fisher transformation (Fisher, 1921) assumes a bivariate normal distribution, so for larger numbers of predictors this value should be taken as a heuristic measure of robustness rather than as an exact statistical value.

cating a strong fit, but with some error variance due to the bootstrap sampling (Table 5).

In summary, Monte Carlo experiments have been utilized to test that the assumed properties (Properties 1–4) of the WRCT (n) measure of robustness hold under situations where data are noisy, and they are robust to error from different error distributions. In addition, the measures can be used as a “heuristic” measure in regression contexts to examine model robustness.

5 Empirical Example

To assess the efficacy of the proposed WCRT method in a survey context, a simple survey was administered to a sample curated through Qualtrics. The goal of the survey was not to investigate any substantive empirical point, but to apply WCRT methods to assess robustness to participant nonresponse bias for a series of correlation tests. As the retailing constructs and scales summarized by Szymanski & Henard (2001) have been widely applied and pose relatively simple questions, they were judged as liable to provide stable results, and unlikely to represent confounding factors due to their complexity.

5.1 The Dataset

The survey includes five different multi-item measurement scales, each of which relates to some measure of customer satisfaction for a recent retail transaction. Each of the individual items is measured using a seven-point Likert scale. The full list of scales and items within these scales is given in Table 6. Overall, there are five different scales, consisting of 19 subitems. The first three scales deal with the actual shopping experience being evaluated, the fourth scale examines how this experience impacts behavioral intent, and the fifth is a general scale measuring retail/shopping enjoyment. Thus, the first three scales should be strongly correlated, while scale five may have some positive correlation with the other scales (someone who has positive views of retail shopping is more likely to select a positive shopping experience), but the level of correlation should be lower. Two of the items (item two on INTENT and item one on ENJOY) were negative direction items and were reversed.

Table 5

OLS Regression on Monte Carlo Simulation

			95% C.I.	
IV	*Coef.*	*Std. Err.*	*Upper*	*Lower*
SystolicBP(P)	109.9***	2.08	105.8	114.0
DiastolicBP(P)	24.0***	2.07	20.0	28.1
Gender(P)	90.7***	2.07	86.6	94.8
Age(P)	786.0***	2.07	781.9	790.1
Income(P)	79.5***	2.07	75.4	83.6
Ht(P)	85.1***	2.07	81.1	89.2
Wt71(P)	237.6***	2.07	233.5	241.6
Wt82(P)	95.6***	2.07	91.5	99.7
SmokeIntensity(P)	19.5***	2.07	15.4	23.6
SmokeYears(P)	355.1***	2.07	359.2	359.2
AlcoholFreq(P)	66.0***	2.07	62.0	70.1
Exercise(P)	30.4***	2.07	26.4	34.5
WCRTEffectMedium	–1924.4***	2.54	–1929.4	–1919.4
WCRTEffectLarge	–2367.5***	2.54	–2372.5	–2362.6
Alpha0.05	183.2***	2.54	178.2	188.2
Alpha0.1	279.6***	2.54	274.6	284.6
Intercept	1833.5***	4.29	1825.1	1842.0
Observations			183,512	
R^2			0.87	

*$p<0.05$, **$p<0.01$, ***$p<0.001$

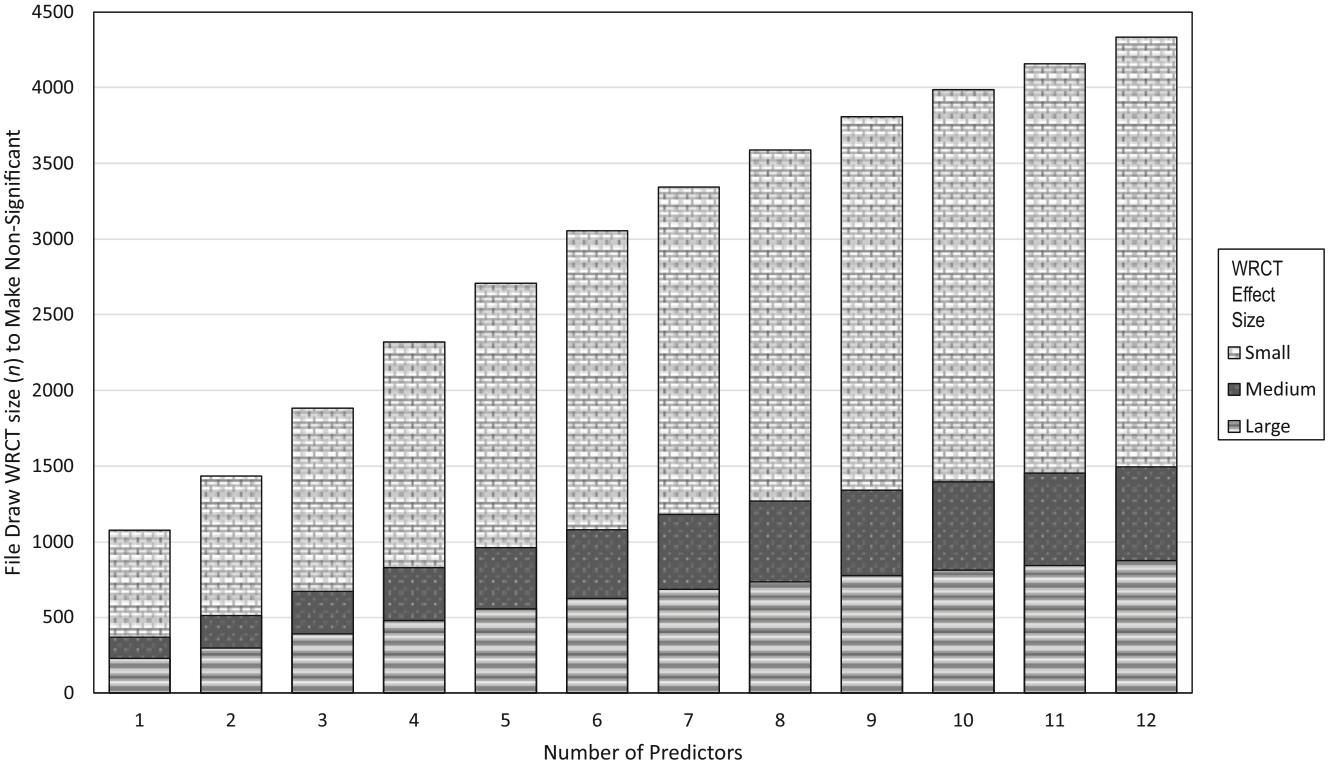

Fig. 4

The File Drawer Number of Studies to Negate Significance For Regression on Health Data Across Number of Predictors and Alpha (α)

The data were collected via a Qualtrics panel. There were n = 415 fully completed surveys out of a total of n = 463 surveys. In line with the focus on participant nonresponse bias, participant responses with missing items were removed rather than imputed using a missing data technique.

5.2 Exploratory Data Analysis

As a preface to analyzing the correlation tests using WCRT, some analysis was performed on the consistency of the rating scale and on the correlations. To examine the consistency of the summated ratings scales, Cronbach's alpha (Cronbach, 1951), was calculated for each of the summated rating scales. The values are EXP (0.96), SAT (0.99), PWM (0.96), INTENT (0.78), and ENJOY (0.78). From past literature (e.g., Bland & Altman, 1997; Tavakol & Dennick, 2011), cut-offs for "good" to "excellent" values of alpha range from 0.7–0.95, so these values are in the correct range.

A summary matrix plot of the overall correlations between the values in the summated rating scales is given in Fig. 5. Here, the diagonal values give histogram distributions of the summated values, the upper triangle of the matrix contains the correlations between the summated values (*** represents $p < 0.001$ for a statistical test of correlation), and the lower triangle contains scatterplots, each overlaid with a linear regression best fit line and a confidence circle for the multivariate mean of the distribution.

5.3 Wave Analysis

A simple wave analysis was performed on the data. For this experiment, respondents were taken from a panel. As noted previously, there are n = 415 fully completed survey forms out of n = 463. For a panel, it is difficult to estimate the number of missing responses, but it is possible to estimate the percentage of missing participants given reported percentages for previous similar studies in the literature. For the purpose of this analysis, three scenarios were assumed, one with 50% response, one with 25% response, and one with 10% response.

The wave analysis approach described in Armstrong and Overton (1977), considers two different waves of responses, an early wave and a late wave, and then a "virtual" wave of nonresponses. While the responses to the survey were not

Table 6

Survey Scale Information

Information	*Description*
Name	Shopping Experience (EXP)
Prompt	Thinking about this retail shopping experience, please rate your overall feelings about the shopping experience
Sub-items	Unpleasant:pleasant dislike very much:like very much left me feeling bad:left me with a good feeling
Name	Satisfaction (SAT)
Prompt	My overall impression of this retail shopping experience is
Sub-items	Bad:Good Unfavorable:Favorable Unsatisfactory:Satisfactory Negative:Positive Dislike:Liked
Name	Positive Word of Mouth (PWOM)
Prompt	Thinking about your shopping experience, please rate your agreement with the following statements
Sub-items	(All strongly disagree:strongly agree) I would say positive things about this retailer. I would recommend this retailer to people I know. I would encourage relatives and friends to do business with this retailer.
Name	Behavioral Intentions (INTENT)
Prompt	Thinking about your shopping experience, please rate your agreement with the following statements
Sub-items	(All strongly disagree:strongly agree) I expect to be coming to this retailer for a long time. I do not expect to visit this retailer in the future. I expect my relationship with this retailer to be enduring. It is likely that I will visit this retailer in the future.
Name	Shopping Enjoyment (ENJOY)
Prompt	Please rate your agreement with the following statements
Sub-items	(All strongly disagree:strongly agree) I consider shopping a big hassle. When traveling, I enjoy visiting new and interesting shops. I enjoy browsing for things even if I cannot buy them yet. I often visit shopping malls or markets just for something to do.

split into waves, for the purpose of this illustrative example, it was assumed that the first 50% belong to the early wave and the second 50% belong to the late wave. The three response scenarios give the number of participant responses for the third wave as 415 (50% response), 1245 (25% response), and 3735 (10% response). Armstrong and Overton (1977) give three methods for calculating values for the third (nonresponse) wave. These are adapted into an effect size context below.

1. Assume that the nonresponses have the same effect size as the second wave.
2. Assume that the nonresponses have the same effect size as the responses at the end of the second wave.
3. Assume a linear interpolation through the nonresponse third wave.

For the measure of interest, let the effect sizes for waves one and two respectively be φ_1 and φ_2. The number of item values in the three waves are denoted n_1, n_2, and n_3. Wave analysis aims to give a prediction for φ_3 in the nonresponse wave. For method one, $\varphi_3 = \varphi_2$. Methods two and three assume a linear relationship for the effect size over time.

[5] To be consistent with the development of the WCRT method, the calculations are given using group means rather than upper and lower boundaries, but the calculations are equivalent.

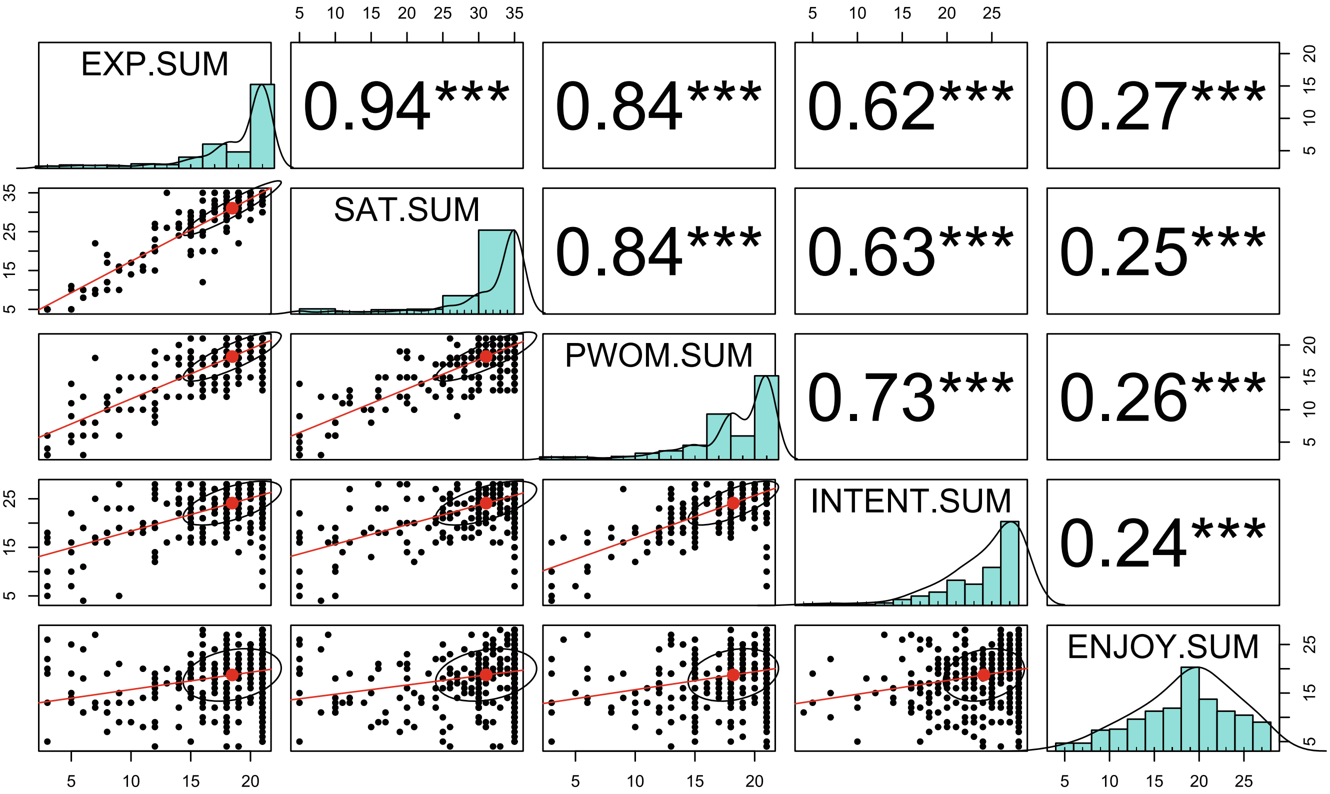

Fig. 5

Multi-Item Scale Correlations

From Armstrong and Overton (1977)[5], for method two, φ_3 is calculated as in (16).

$$\varphi_3 = \varphi_2 + (\varphi_2 - \varphi_1)\,\frac{n_2}{(n_1 + n_2)} \tag{16}$$

Here, a straight line is drawn between the midpoint of group one and the midpoint of group two. The line is extrapolated to the end of group two. For method three, the line is extrapolated to the middle of group three, giving (17).

$$\varphi_3 = \varphi_2 + (\varphi_2 - \varphi_1)\,\frac{(n_2 + n_3)}{(n_1 + n_2)} \tag{17}$$

A wave analysis was performed for each combination of the correlations given in Fig. 5 and the three different nonresponse scenarios. Here, the correlation r is the effect size. The first two wave analysis methods are independent of the number of nonresponses n_3, but the third is not, so the results differ across the three nonresponse scenarios.

The results of the wave analysis are given in Table 7. Results are given for each of the 10 possible correlations between the summated rating scales. The first two columns contain the values of the mean correlation values for waves one and two. The means for the second wave are taken as the M1 (method one) estimate of the third wave. The next column contains the M2 (method two) estimates of the correlations at the end of the second wave and the subsequent columns contain the M3 (method three) estimates for the three levels of participant response (50%, 25%, and 10%). For the moderate response scenarios (50%, 25%), the correlations all stayed within bounds, but for the 10% scenario, several values needed to be truncated at either –1 or 1. This shows the difficulty of a linear interpolation that extends well beyond the range of data. It is likely that as n_3 increases, any change in the dependent variable will lessen. However, the values for 10% response provide useful "extreme bounds", which can be utilized by the WCRT procedure.

5.4 WCRT Procedure

As previously shown in Fig. 5, all multi-item scale correlations are strongly ($p < 0.001$) significant, with correlations between scales related to the actual shopping experience (EXP, SAT, POW) over 0.8, correlations between these scales and the future shopping intention (INTENT)

Table 7

Wave Analysis Results

Correlation	r_1	*M1:* r_2	*M2: End wave 2*	*M3: 415 (50%)*	*M3: 1245 (25%)*	*M3: 3735 (10%)*
EXP, SAT	0.93	0.96	0.97	1.00	1.00	1.00
EXP, PWOM	0.86	0.82	0.80	0.75	0.67	0.42
EXP, INTENT	0.71	0.52	0.42	0.23	–0.15	–1.00
EXP, ENJOY	0.31	0.22	0.18	0.09	–0.08	–0.61
SAT, PWOM	0.88	0.80	0.76	0.67	0.50	0.00
SAT, INTENT	0.74	0.50	0.38	0.14	–0.33	–1.00
SAT, ENJOY	0.29	0.20	0.16	0.08	–0.09	–0.58
PWOM, INTENT	0.81	0.63	0.53	0.34	–0.04	–1.00
PWOM, ENJOY	0.27	0.25	0.24	0.22	0.19	0.09
INTENT, ENJOY	0.26	0.21	0.18	0.13	0.02	–0.30

scale in the 0.6–0.7 range and the correlations between the general shopping enjoyment measure and the other scales in the 0.2–0.3 range.

For each correlation, the WCRT procedure was calculated for opposing effect sizes with increments of 0.01 ranging from –0.99 to the maximum effect size with a finite n (approximately 0) for alpha (α) values of 0.01, 0.05, and 0.1. Selected results are examined in Figs. 6 and 7 in what we call "n-curves", which are similar to the n-curves that have been used to determine sample sizes (e.g., Trafimow, 2018) and the probability of replication (Killeen, 2005), and the previously discussed p-curves for statistical power (Simonsohn et al., 2014).

For contrast, curves are given for the highest correlation (EXT and SAT), where $r_1 = 0.94$, and for the lowest correlation (INTENT and JOY), where $r_1 = 0.24$. For each of these correlations, curves are given for $\alpha = 0.05$, though any value of α can be chosen. The x-axis contains the r_2 required to negate the significance of the significance test[6]. In the case of correlations, due to the asymptotic behavior of the significance test being a tradeoff between the overall effect size and n, only negative r_2 values give a finite n and the graphs go off to infinity at approximately $r_2 = 0$.

As the relationship between the value of r and n is strongly exponential, it is difficult to plot n versus r on a linear scale, so a logarithmic scale is used for n. This makes it more difficult to read the values of n, but to make up for this, values of n are explicitly given for negatives of the standard effect sizes defined by Cohen (1988), giving $r = –0.1$ (small), $r = –0.3$ (medium) and $r = –0.5$ (large) effect sizes, along with $r = –0.7$ and $r = –0.9$.

Looking at Fig. 6, which is for an $\alpha = 0.05$ test for the pair of scales with the strongest correlation ($r_1 = 0.94$), for a small negative effect ($r_2 = –0.1$), $n = 5670$ would be required to negate significance, while for a large negative effect ($r_2 = –0.5$), $n = 1175$ would be required to negate significance. This would be very unlikely, given the large negative effect. Even an almost "complete reversal" of the correlation ($r_2 = –0.9$) would require $n = 454$ in order to negate significance.

The graph in Fig. 7 is for an α = 0.05 test for the lowest correlation of $r = 0.24$ between INTENT and ENJOY, and it shows much lower values of n. For $\alpha = 0.05$, for a small negative effect ($r_2 = –0.1$), $n = 427$ would be required to negate significance, while for a large negative effect ($r_2 = –0.5$), $n = 103$ would be required to negate significance. The extreme $r_2 = –0.9$ case would require $n = 43$ to negate significance.

5.5 Combining WCRT with Wave Analysis

In any scenario where response times can be calculated, wave analysis can provide estimates of sample statistics for nonresponding participants, which can be converted to effect sizes. These effect sizes can be used to help choose a realistic range of effect sizes in the outlined WCRT procedure. Accordingly, we propose a method combining wave analysis results with WCRT to create a set of "warning" metrics for results that may be called in to question by possible nonresponse bias. An outline of the method is given below.

Assume a situation where a statistical test has been performed with some level of Type I error α and there are two possible results; either H_0 is rejected in favor of H_A or there is not enough evidence to reject H_0. The test will have some

[6] Similar n-curves could be drawn where the aim is to find n to make a non-significant test significant.

measure of effect size (e.g., Cohen's d for a two-sample test or the sample correlation r for a correlation test). There is some number of nonresponses n_3.

1. Calculate the three different wave analysis effect size values: M1: average of second wave, M2: end of second wave, M3: extrapolation to mean of third (nonresponse) wave.
2. For WCRT, calculate the effect size needed to reverse the statistical test given the number of nonresponses n_3. This is the inverse procedure of finding n given an effect size, i.e., for a correlation effect size r, if the calculation of n from r is defined as the function $f(r) = n$ then $f^{-1}(n) = r$.
3. Record if each of the three effect sizes found by wave analysis will reverse the result of the statistical test. For example, for a positive correlation r that is statistically significant, if the correlation predicted by wave analysis for nonrespondents is less (of greater magnitude) than the WCRT r_2 value, then the wave analysis correlation value will reverse the test and the result should be flagged.

The three wave analysis predictions give different levels of future extrapolations. For M1, where the predicted nonresponse effect size is the aggregate effect size for the second wave, unless a statistical result is close to a boundary, it is unlikely that a nonresponse effect size value equal to the value for the second wave will change the result of a statistical test. However, a linear extrapolation for M3 to the middle of the nonresponse wave for large nonresponse n is liable to change a test result and the extrapolation is likely to be over-exaggerated, as it is unlikely that the trend from the first wave to the second wave would continue linearly for a large nonresponse wave. Some damping is likely. However, the M3 scenario can provide a good "worst-case" scenario.

The combined method was applied to the previously discussed correlation example for all 10 correlations, two significance levels ($\alpha = 0.05, 0.01$), and the previously discussed participant nonresponse scenarios (nonresponse n = 415, 1245, 3735). The results are given in Table 8, 9 and 10, with each table containing one of the three nonresponse scenarios. Each table contains a row for each of the ten tested correlations. There are columns for the sample correlation value, the three wave analysis values, and the two WCRT values for the tested values of the Type I error α. As all correlations are significant and positive, the wave analysis results are flagged/counted as reversing the result of the statistical test if the correlations are less than the WCRT values. These flagged correlations are marked with a star (*) symbol.

In Table 8, no values are flagged and none of the wave analysis scenarios will reverse the result of the statistical test. In part, this is because all the test correlations are quite "strong". Even the correlations that include the ENJOY measure ($0.24 \leq r \leq 0.27$), while less than the other correlations, are strongly significant with a sample size of n = 415. As the nonresponse n_3 increases from 415 to 3735, the magnitude of the correlations found by the inverse WCRT procedure decreases. This is intuitive, as given that statistical significance is a function of both effect size and sample size, for a larger sample size, a smaller negative effect is needed to reverse the results of a statistical test.

In Table 9, the extrapolated r_3 for M3 goes outside of the testing "flip" boundaries defined by WCRT for three correlations, which increases to six correlations for the $n_3 = 3735$ results given in Table 10. This includes all the "enjoy" correlations except for the "PWOM, ENJOY" correlation, for which there is only a very slight linear trend. Despite negative linear trends, the "EXP, PWOM" and "SAT, PWOM" correlations are not flagged, as the correlations are high relative to the negative linear trends.

6 Discussion

This study has presented a methodology and set of statistical tools for analyzing nonresponse bias situations. A methodology based on the file draw problem and worst-case resistance testing (WCRT) is given to help researchers quantify and understand the "robustness" of results with respect to nonresponse bias. Researchers can examine the number of nonrespondents needed to reverse the results of a statistical test for a range of feasible effect sizes for the nonresponse data. This relationship can be plotted using an "n-curve". The range of feasible effect sizes can be decided using evidence from past research, guidance on standard effect sizes, or the results of a wave analysis. Conversely, researchers can find the effect size needed to reverse the results of a statistical test for a given number of experimental nonresponses and then evaluate if these effect sizes are feasible using the guidance described above.

The basic WCRT methodology was developed in this paper as a method for analyzing robustness towards nonresponse bias. However, the methodology is more generally applicable to other scenarios. For any situation where there is a statistical test and some idea of possible "negative effect sizes", the WCRT methodology can be used to measure robustness. As noted in the introduction, there is a strong push to improve experimental rigor in the behavioral sciences and in marketing. An added urgency was added to this process by reports finding a low level of replicability in behavioral science studies (e.g., Open Science Collaboration, 2015; Stanley et al., 2018) and by high-profile behavioral research scandals and retractions (e.g., Inman et al., 2018; Stricker & Günther, 2019). In addition to the focus on improving statistical rigor described earlier in the paper (e.g.,

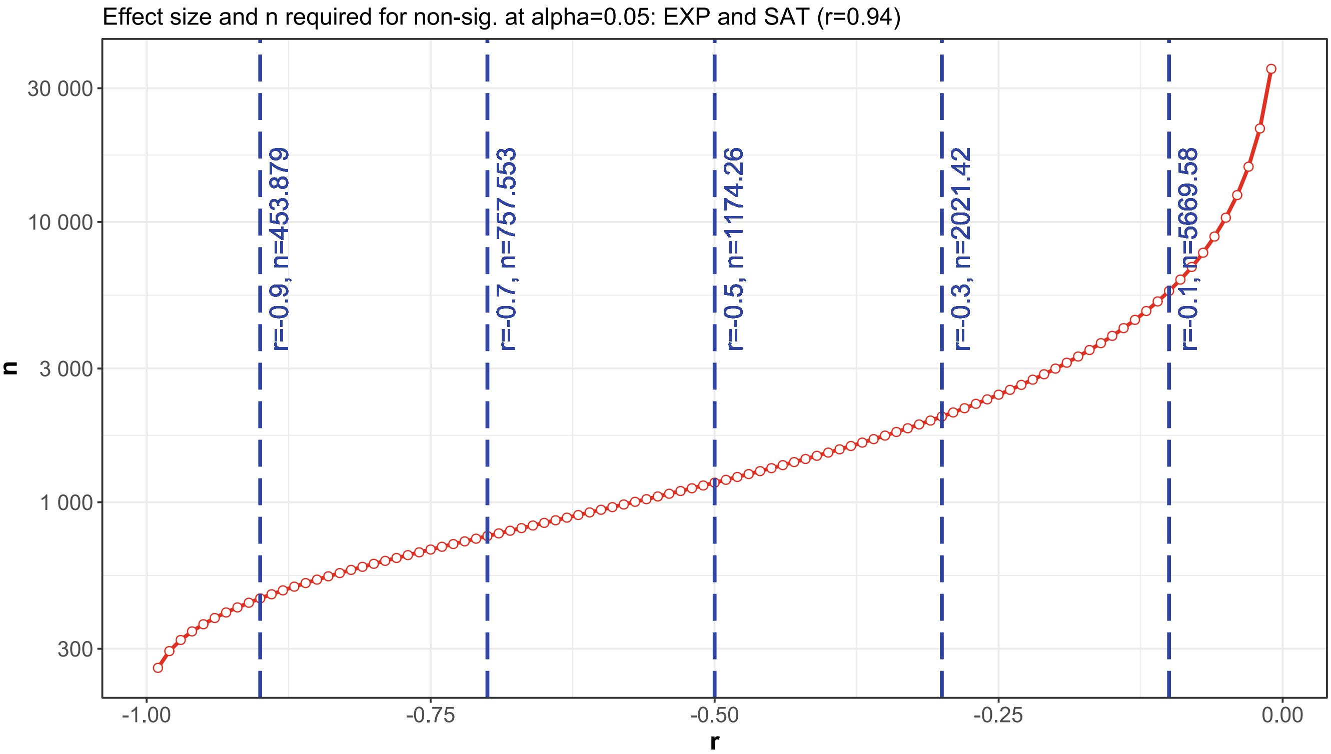

Fig. 6

n-Curve for EXP and SAT: $\alpha = 0.05$

JCR, 2021; Harvey, 2017; Schwab et al., 2011; Wasserstein & Lazar, 2016), there has been a move towards requiring preregistration of experiments (Simmons et al., 2021), i.e., the process of researchers stating the experimental procedure and expected results and storing this information externally in a third-party repository, and to improved sharing and availability of research data (Towse et al., 2021). Including the preregistration information along with a paper submission ensures that the experiment is not altered in an ad-hoc manner to account for unexpected results.

The methods outlined in this paper can easily be incorporated into the behavioral science environment outlined above. Even in a pure experimental setting, some type of nonresponse bias may be present; for example, for a student experiment, a certain number of students in a subject pool could be notified of a study, with only a few participating.

Table 8

Combining Wave Analysis and Worst-case Resistance Testing for 50% Response (Nonresponse n_3 = 415)

Correlation	*r*	*r_3 (M1)*	*r_3 (M2)*	*r_3 (M3)*	*Wr_3 α = 0.05*	*Wr_3 α = 0.01*
EXP, SAT	0.94	0.96	0.97	1.00	–0.93	–0.92
EXP, PWOM	0.84	0.82	0.80	0.75	–0.80	–0.78
EXP, INTENT	0.62	0.52	0.42	0.23	–0.54	–0.51
EXP, ENJOY	0.27	0.22	0.18	0.09	–0.16	–0.11
SAT, PWOM	0.84	0.80	0.76	0.67	–0.80	–0.79
SAT, INTENT	0.63	0.50	0.38	0.14	–0.55	–0.52
SAT, ENJOY	0.25	0.20	0.16	0.08	–0.14	–0.09
PWOM, INTENT	0.73	0.63	0.53	0.34	–0.67	–0.65
PWOM, ENJOY	0.26	0.25	0.24	0.22	–0.15	–0.10
INTENT, ENJOY	0.24	0.21	0.18	0.13	–0.13	–0.08

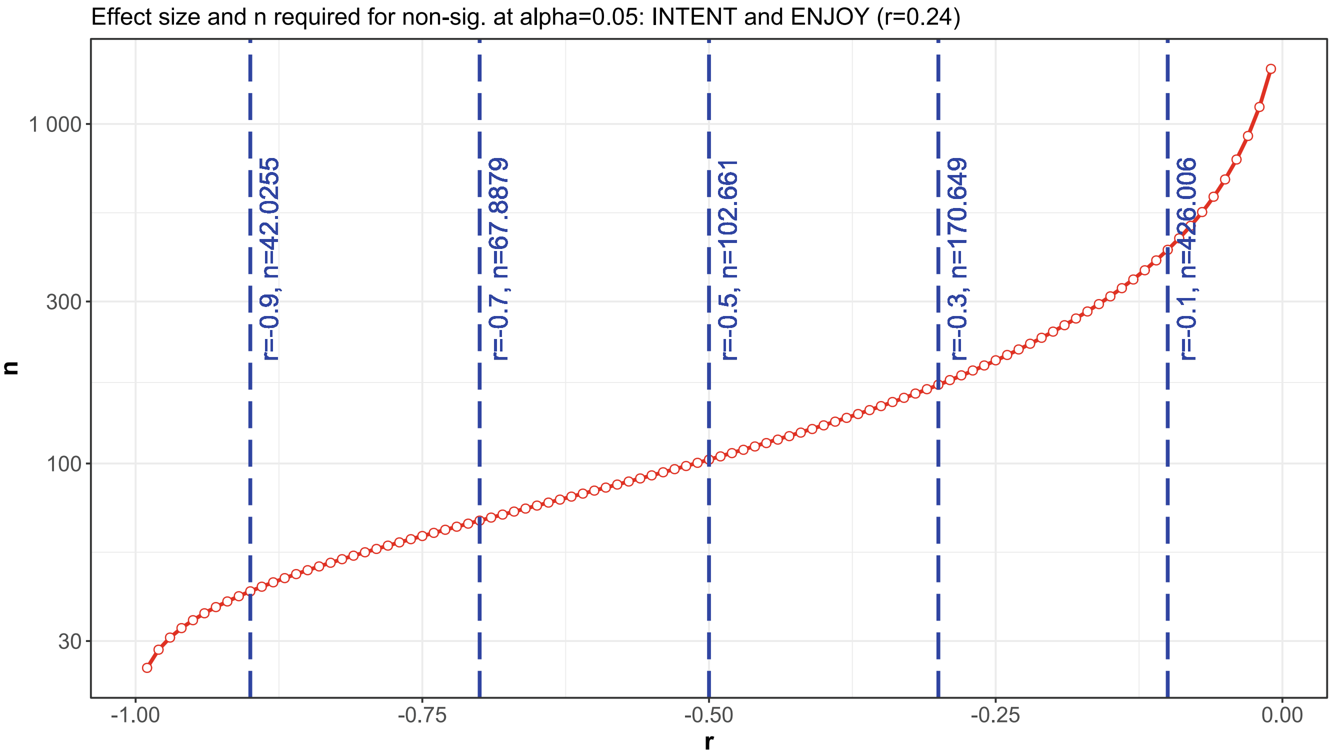

Fig. 7

n-Curve for INTENT and ENJOY: α = 0.05

Table 9

Combining Wave Analysis and Worst-Case Resistance Testing for 25% Response (Nonresponse n_3 = 1245)

Correlation	*r*	*r_3 (M1)*	*r_3 (M2)*	*r_3 (M3)*	*Wr_3 α = 0.05*	*Wr_3 α = 0.01*
EXP, SAT	0.94	0.96	0.97	1.00	–0.48	–0.47
EXP, PWOM	0.84	0.82	0.80	0.67	–0.34	–0.32
EXP, INTENT	0.62	0.52	0.42	–0.15	–0.18	–0.16
EXP, ENJOY	0.27	0.22	0.18	–0.08	–0.02*	–0.00*
SAT, PWOM	0.84	0.80	0.76	0.50	–0.34	–0.32
SAT, INTENT	0.63	0.50	0.38	–0.33	–0.19*	–0.17*
SAT, ENJOY	0.25	0.20	0.16	–0.09	–0.03*	0.01*
PWOM, INTENT	0.73	0.63	0.53	–0.04	–0.25	–0.23
PWOM, ENJOY	0.26	0.25	0.24	0.19	–0.03	–0.01
INTENT, ENJOY	0.24	0.21	0.18	0.02	–0.03	0.00

When nonresponse bias is not an issue, WCRT can still be used to help examine the robustness of the results. Gelman and Loken (2013) noted that even with preregistration and no p-hacking, researchers can still bend the rules, for example, choosing the regression technique that gives the best results or choosing whether to use a main effect or interaction effect to justify a hypothesis. Given continued publication bias towards significant results (e.g., Franco et al. 2014; Harrison et al. 2017), there will always be an incentive to choose the research path to give the most significant results, in what statisticians sometimes call "the garden of forking paths". Rules to increase experimental rigor, such as preregistration, may prune some of these paths, but without being overly restrictive, cannot prevent researchers finding new paths. This is somewhat analogous to the situation of

Table 10

Combining Wave Analysis and Worst-Case Resistance Testing for 10% Response (Nonresponse n_3 = 3735)

Correlation	*r*	*r_3 (M1)*	*r_3 (M2)*	*r_3 (M3)*	*Wr_3 α = 0.05*	*Wr_3 α = 0.01*
EXP, SAT	0.94	0.96	0.97	1.00	–0.16	–0.15
EXP, PWOM	0.84	0.82	0.80	0.42	–0.11	–0.09
EXP, INTENT	0.62	0.52	0.42	–1.00	–0.05*	–0.04*
EXP, ENJOY	0.27	0.22	0.18	–0.61	0.00*	0.01*
SAT, PWOM	0.84	0.80	0.76	0.00	–0.11	–0.09
SAT, INTENT	0.63	0.50	0.38	–1.00	–0.05*	–0.04*
SAT, ENJOY	0.25	0.20	0.16	–0.58	0.01*	0.02*
PWOM, INTENT	0.73	0.63	0.53	–1.00	–0.07*	–0.06*
PWOM, ENJOY	0.26	0.25	0.24	0.09	0.00	0.01
INTENT, ENJOY	0.24	0.21	0.18	–0.30	0.00*	0.01*

accountants finding new workarounds as rules on tax avoidance are strengthened.

In the context outlined above, WCRT could be utilized as a measure of robustness of results with respect to all possible experimental errors and biases. A range of possible effect sizes for the nonresponse bias could be derived and combined. Feasible nonresponse effect sizes could be derived for nonresponses using wave analysis or using any method for creating feasible bounds (for example, Manski bounds), by collating effect sizes the past literature in the area, or through a meta-analytic p-curve analysis (Simonsohn et al. 2014). In time, a set of "*n*" thresholds could be developed to flag results with insufficient robustness to the factors outlined above.

6.1 Limitations and Future Research

This paper develops WCRT methods for hypothesis tests of means, correlations, and simple regression scenarios. To be widely utilized, WCRT methods would need to be developed for a wider range of statistical tests, such as GLM (general linear models) and SEM (structural equation models), as these methods are the most widely used methods in behavioral research. For example, the second Monte Carlo experiment shows how WRCT methods can be utilized to create a heuristic measure of robustness for regression. However, exact statistical inference is limited by the assumptions of the z transform (Fisher, 1921). In meta-analysis regression, multiple groups for different studies are often handled using multi-group fixed-effect or random-effects regression. There is scope to apply these methods (e.g., Borenstein et al. 2010; Hedges & Vevea, 1998) to model WCRT response vs. nonresponse and to create a generalized methodology for WCRT regression analysis.

The scenario outlined above is similar to the scenario that has unfolded in the area of effect size and power calculations, where over time, methods have been developed for a wide range of statistical tests. For the WCRT methods described in this paper to be widely used, it would be important to package them together into a single cohesive software package, in a similar manner to G*Power (Faul et al., 2007), which has become the de-facto standard software package for power analysis.

In the modern internet-mediated environment, more surveys are being conducted using online panels designed to represent certain population characteristics and through coworking/online hiring platforms, such as the Amazon Mechanical Turk (Kees et al., 2017). Determining nonresponse in online environments is difficult, as the survey platform recruitment procedure may be opaque. What exactly constitutes nonresponse in a panel or online working platform? If a set of respondents are notified about an opportunity, then the number of nonresponses can be calculated only if the number notified is reported by the platform. In a coworking platform where respondents search through lists of opportunities, calculating nonresponse may be difficult. If views of an opportunity are recorded (e.g., through a scroll-down list), then some measure of nonresponse of "aware" respondents can be calculated, but determining how to set a threshold for awareness would be difficult. There has been some initial work on analyzing nonresponse for the Mechanical Turk for longitudinal studies (Daly & Nataraajan, 2015) and several studies have tried to quantify possible nonresponse bias for online platforms (Boas et al., 2020; Paolacci et al., 2010). However, there is strong scope for a systematic analysis of nonresponse for online surveys. Such analysis could include work from both information systems and experimental standpoints, and could include aspects such as data reporting, human-computer interaction, and nonresponse behavior.

The wave analysis method utilized in this paper is a simple linear extrapolation method. Linear extrapolation may not be reliable outside of the range of the data. It is likely that significant linear trends would probably "damp" outside of the range of the data, particularly in situations where there are many nonrespondents. This is a reason why damped trend forecasting methods that give conservative forecasts are often successful (e.g., Armstrong et al., 2015; Gardner, 2015). For the use of wave analysis in the experimental section, this lack of conservatism is an advantage, as linear extrapolation is used to create worst case bounds for correlations. However, given the advances in forecasting over the 40 plus years since the introduction of wave analysis (e.g., Makridakis et al., 2020), there is scope to bring new methodology to bear on wave analysis and develop methods to improve forecasts of nonresponse bias.

Open Practices Statement The dataset collated for this paper and the code for the procedures developed in the paper have made available at https://github.com/MDSOPT/WCRT.

References

Aiken, L.R. (1981). Proportion of returns in survey research. *Educational and Psychological Measurement*, *41*(4), 1033–1038.

Akkucuk, U., & Carroll, J.D. (2006). PARAMAP vs. Isomap: a comparison of two nonlinear mapping algorithms. *Journal of Classification*, *23*, 221–254.

Armstrong, J.S., & Overton, T.S. (1977). Estimating nonresponse bias in mail surveys. *Journal of Marketing Research*, *14*(3), 396–402.

Armstrong, J.S., Green, K.C., & Graefe, A. (2015). Golden rule of forecasting: be conservative. *Journal of Business Research*, *68*(8), 1717–1731.

Babić Rosario, A., Sotgiu, F., De Valck, K., & Bijmolt, T.H. (2016). The effect of electronic word of mouth on sales: a meta-analytic review of platform, product, and metric factors. *Journal of Marketing Research*, *53*(3), 297–318.

Banfield, J.D., & Raftery, A.E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics*, *49*(3), 803–821.

Baroudi, J.J., & Orlikowski, W.J. (1989). The problem of statistical power in MIS research. *MIS Quarterly*, *13*(1), 87–106.

Beebe, T.J., Talley, N.J., Camilleri, M., Jenkins, S.M., Anderson, K.J., & Locke III, G.R. (2011). Health insurance portability and accountability act (HIPAA) authorization and survey nonresponse bias. *Medical Care*, *49*(4), 365–370.

Berg, N. (2005). Non-response bias. In K. Kempf-Leonard (Ed.), *Encyclopedia of social measurement* (Vol. 2, pp. 865–873). Academic Press.

Bland, J.M., & Altman, D.G. (1997). Statistics notes: Cronbach's alpha. *BMJ*, *314*(7080), 572.

Boas, T.C., Christenson, D.P., & Glick, D.M. (2020). Recruiting large online samples in the United States and India: Facebook, Mechanical Turk, and Qualtrics. *Political Science Research and Methods*, *8*(2), 232–250.

Borenstein, M., Hedges, L.V., Higgins, J.P., & Rothstein, H.R. (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods*, *1*(2), 97–111.

Boyd Jr., H.W., & Westfall, R. (1965). Interviewer bias revisited. *Journal of Marketing Research*, *2*(1), 58–63.

Brusco, M.J., Singh, R., Cradit, J.D., & Steinley, D. (2017). Cluster analysis in empirical OM research: survey and recommendations. *International Journal of Operations & Production Management*, *37*(3), 300–320.

Cashen, L.H., & Geiger, S.W. (2004). Statistical power and the testing of null hypotheses: a review of contemporary management research and recommendations for future studies. *Organizational Research Methods*, *7*(2), 151–167.

Chandler, J., Rosenzweig, C., Moss, A.J., Robinson, J., & Litman, L. (2019). Online panels in social science research: expanding sampling methods beyond Mechanical Turk. *Behavior Research Methods*, *51*(5), 2022–2038.

Chen, L., & Buja, A. (2009). Local multidimensional scaling for nonlinear dimension reduction, graph drawing, and proximity analysis. *Journal of the American Statistical Association*, *104*(485), 209–219.

Chesney, D.L., & Obrecht, N.A. (2012). Statistical judgments are influenced by the implied likelihood that samples represent the same population. *Memory & Cognition*, *40*(3), 420–433.

Coe, R. (2002). It's the effect size, stupid: what effect size is and why it is important. In *Annual Conference of the British Educational Research Association* (pp. 1–18). British Educational Research Association.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd edn.). Lawrence Erlbaum.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155–159.

Collier, J.E., & Bienstock, C.C. (2007). An analysis of how nonresponse error is assessed in academic marketing research. *Marketing Theory*, *7*(2), 163–183.

Cook, C., Heath, F., & Thompson, R.L. (2000). A meta-analysis of response rates in web-or internet-based surveys. *Educational and Psychological Measurement*, *60*(6), 821–836.

Cox, N.J. (2008). Speaking Stata: Correlation with confidence, or Fisher's z revisited. *The Stata Journal*, *8*(3), 413–439.

Cronbach, L.J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334.

Daikeler, J., Bošnjak, M., & Lozar Manfreda, K. (2020). Web versus other survey modes: an updated and extended meta-analysis comparing response rates. *Journal of Survey Statistics and Methodology*, *8*(3), 513–539.

Daly, T.M., & Nataraajan, R. (2015). Swapping bricks for clicks: crowdsourcing longitudinal data on Amazon turk. *Journal of Business Research*, *68*(12), 2603–2609.

Daniel, W.W., Schott, B., Atkins, F.C., & Davis, A. (1982). An adjustment for nonresponse in sample surveys. *Educational and Psychological Measurement*, *42*(1), 57–67.

Deming, W.E. (1953). On a probability mechanism to attain an economic balance between the resultant error of response and the bias of nonresponse. *Journal of the American Statistical Association*, *48*(264), 743–772.

Diamantopoulos, A., & Winklhofer, H.M. (2001). Index construction with formative indicators: an alternative to scale development. *Journal of Marketing Research*, *38*(2), 269–277.

Faul, F., Erdfelder, E., Lang, A.G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191.

Field, A.P. (2001). Meta-analysis of correlation coefficients: A Monte Carlo comparison of fixed- and random-effects methods. *Psychological Methods*, *6*(2), 161–180.

Fisher, R.A. (1921). On the "probable error" of a coefficient of correlation deduced from a small sample. *Metron*, *1*, 13–32.

France, S.L., Adams, F.G., & Landers, V.M. (2024a). *Dataset for worst case resistance testing: a nonresponse bias solution for today's survey research realities*. Ann Arbor: Inter-university Consortium for Political and Social Research [distributor]. https://doi.org/10.3886/E203261V2.

France, S.L., Adams, F., & Landers, M. (2024b). Software for worst case resistance testing: nonresponse bias solution for today's survey research realities. https://github.com/MDSOPT/WCRT

Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: unlocking the file drawer. *Science*, *345*(6203), 1502–1505.

Gardner, E.S. (2015). Conservative forecasting with the damped trend. *Journal of Business Research*, *68*(8), 1739–1741.

Gelman, A., & Loken, E. (2013). The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time. https://stat.columbia.edu/~gelman/research/unpublished/forking.pdf

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

Groves, R.M. (2006). Nonresponse rates and nonresponse bias in household surveys. *Public Opinion Quarterly*, *70*(5), 646–675.

Groves, R.M., & Peytcheva, E. (2008). The impact of nonresponse rates on nonresponse bias: a meta-analysis. *Public Opinion Quarterly*, *72*(2), 167–189.

Halbesleben, J.R.B., & Whitman, M.V. (2013). Evaluating survey quality in health services research: a decision framework for assessing nonresponse bias. *Health Services Research*, *48*(3), 913–930.

Harrison, J.S., Banks, G.C., Pollack, J.M., O'Boyle, E.H., & Short, J. (2017). Publication bias in strategic management research. *Journal of Management*, *43*(2), 400–425.

Hartman, B.W., Fuqua, D.R., & Jenkins, S.J. (1986). The problems of and remedies for nonresponse bias in educational surveys. *The Journal of Experimental Education*, *54*(2), 85–90.

Harvey, C.R. (2017). Presidential address: the scientific outlook in financial economics. *The Journal of Finance*, *72*(4), 1399–1440.

Hedges, L.V., & Vevea, J.L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods*, *3*(4), 486.

Hemphill, J.F. (2003). Interpreting the magnitudes of correlation coefficients. *American Psychologist*, *58*(1), 78–79.

Hernán, M.A., & Robins, J.M. (2020). Causal inference. Chapman & Hall/CRC. https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/

Higgins, J.P., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M.J., & Welch, V.A. (Eds.). (2019). *Cochrane handbook for systematic reviews of interventions*. John Wiley & Sons.

Horowitz, J.L., & Manski, C.F. (1998). Censoring of outcomes and regressors due to survey nonresponse: identification and estimation using weights and imputations. *Journal of Econometrics*, *84*(1), 37–58.

Hubbard, R., & Armstrong, J.S. (2006). Why we don't really know what statistical significance means: implications for educators. *Journal of Marketing Education*, *28*(2), 114–120.

Hunter, J.E. (1997). Needed: a ban on the significance test. *Psychological Science*, *8*(1), 3–7.

Inman, J.J., Campbell, M.C., Kirmani, A., & Price, L.L. (2018). Our vision for the Journal of Consumer Research: it's all about the consumer. *Journal of Consumer Research*, *44*(5), 955–959.

JCR (2021). Journal of Consumer Research: Research Ethics. https://consumerresearcher.com/research-ethics. Accessed 02.06.

Kanuk, L., & Berenson, C. (1975). Mail surveys and response rates: a literature review. *Journal of Marketing Research*, *12*(4), 440–453.

Kees, J., Berry, C., Burton, S., & Sheehan, K. (2017). An analysis of data quality: professional panels, student subject pools, and Amazon's Mechanical Turk. *Journal of Advertising*, *46*(1), 141–155.

Kelley, K., & Preacher, K.J. (2012). On effect size. *Psychological Methods*, *17*(2), 137–152.

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

Lambert, D.M., & Harrington, T.C. (1990). Measuring nonresponse bias in customer service mail surveys. *Journal of Business Logistics*, *11*(2), 5–25.

Lee, S. (2006). Propensity score adjustment as a weighting scheme for volunteer panel web surveys. *Journal of Official Statistics*, *22*(2), 329.

Leventhal, L., & Huynh, C.L. (1996). Directional decisions for two-tailed tests: Power, error rates, and sample size. *Psychological Methods*, *1*(3), 278.

MacDonald, S.E., Newburn-Cook, C.V., Schopflocher, D., & Richter, S. (2009). Addressing nonresponse bias in postal surveys. *Public Health Nursing*, *26*(1), 95–105.

Makridakis, S., Hyndman, R.J., & Petropoulos, F. (2020). Forecasting in social settings: the state of the art. *International Journal of Forecasting*, *36*(1), 15–28.

Manski, C.F. (2016). Credible interval estimates for official statistics with survey nonresponse. *Journal of Econometrics*, *191*(2), 293–301.

Mende, M., Scott, M.L., van Doorn, J., Grewal, D., & Shanks, I. (2019). Service robots rising: how humanoid robots influence service experiences and elicit compensatory consumer responses. *Journal of Marketing Research*, *56*(4), 535–556.

Milligan, G.W. (1981). A Monte Carlo study of thirty internal criterion measures for cluster analysis. *Psychometrika*, *46*, 187–199.

Mohr, J., & Spekman, R. (1994). Characteristics of partnership success: partnership attributes, communication behavior, and conflict resolution techniques. *Strategic Management Journal*, *15*(2), 135–152.

Newman, D.A. (2009). Missing data techniques and low response rates. In I.C.E. Lance & R.J. Vandenberg (Eds.), *Statistical and methodological myths and Urban legends: doctrine, verity and fable in the organizational and social sciences* (pp. 7–36). Routledge.

Olson, K. (2006). Survey participation, nonresponse bias, measurement error bias, and total bias. *Public Opinion Quarterly*, *70*(5), 737–758.

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716-1–aac4716-8.

Paolacci, G., Chandler, J., & Ipeirotis, P.G. (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making*, *5*(5), 411–419.

Pavlou, P.A. (2003). Consumer acceptance of electronic commerce: integrating trust and risk with the technology acceptance model. *International Journal of Electronic Commerce*, *7*(3), 101–134.

Peer, E., Brandimarte, L., Samat, S., & Acquisti, A. (2017). Beyond the Turk: alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*, *70*, 153–163.

Qualtrics (2020). Online samples. https://www.qualtrics.com/research-services/online-sample/. Accessed 04.12.

Rogelberg, S.G., & Stanton, J.M. (2007). Introduction: understanding and dealing with organizational survey nonresponse. *Organizational Research Methods*, *10*(2), 195–209.

Rogelberg, S.G., Luong, A., Sederburg, M.E., & Cristol, D.S. (2000). Employee attitude surveys: examining the attitudes of noncompliant employees. *Journal of Applied Psychology*, *85*(2), 284–293.

Rogelberg, S.G., Conway, J.M., Sederburg, M.E., Spitzmüller, C., Aziz, S., & Knight, W.E. (2003). Profiling active and passive nonrespondents to an organizational survey. *Journal of Applied Psychology*, *88*(6), 1104.

Rosenberg, M.S. (2005). The file-drawer problem revisited: a general weighted method for calculating fail-safe numbers in meta-analysis. *Evolution*, *59*(2), 464–468.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*(3), 638–641.

Rosenthal, R., & Rubin, D.B. (1982). Comparing effect sizes of independent studies. *Psychological Bulletin*, *92*(2), 500–504.

Roth, P.L. (1994). Missing data: a conceptual review for applied psychologists. *Personnel Psychology*, *47*(3), 537–560.

Rotnitzky, A., Robins, J.M., & Scharfstein, D.O. (1998). Semiparametric regression for repeated outcomes with nonignorable nonresponse. *Journal of the American Statistical Association*, *93*(444), 1321–1339.

Sawyer, A.G., & Ball, A.D. (1981). Statistical power and effect size in marketing research. *Journal of Marketing Research*, *18*(3), 275–290.

Scharfstein, D.O., & Irizarry, R.A. (2003). Generalized additive selection models for the analysis of studies with potentially nonignorable missing outcome data. *Biometrics*, *59*(3), 601–613.

Schneider, J.W. (2015). Null hypothesis significance tests. A mix-up of two different theories: the basis for widespread confusion and numerous misinterpretations. *Scientometrics*, *102*(1), 411–432.

Schonlau, M., Van Soest, A., Kapteyn, A., & Couper, M. (2009). Selection bias in web surveys and the use of propensity scores. *Sociological Methods & Research*, *37*(3), 291–318.

Schwab, A., Abrahamson, E., Starbuck, W.H., & Fidler, F. (2011). Researchers should make thoughtful assessments instead of null-hypothesis significance tests. *Organization Science*, *22*(4), 1105–1120.

Sevilla, J., & Townsend, C. (2016). The space-to-product ratio effect: How interstitial space influences product aesthetic appeal, store perceptions, and product preference. *Journal of Marketing Research*, *53*(5), 665–681.

Simmons, J.P., Nelson, L.D., & Simonsohn, U. (2011). False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366.

Simmons, J.P., Nelson, L.D., & Simonsohn, U. (2021). Pre-registration: why and how. *Journal of Consumer Psychology*, *31*(1), 151–162.

Simonsohn, U., Nelson, L.D., & Simmons, J.P. (2014). P-curve: a key to the file-drawer. *Journal of Experimental Psychology: General*, *143*(2), 534.

Sirdeshmukh, D., Singh, J., & Sabol, B. (2002). Consumer trust, value, and loyalty in relational exchanges. *Journal of Marketing*, *66*(1), 15–37.

Sitter, R.R. (1992). Comparing three bootstrap methods for survey data. *Canadian Journal of Statistics*, *20*(2), 135–154.

Skafida, V., Morrison, F., & Devaney, J. (2022). Answer refused: exploring how item non-response on domestic abuse questions in a social survey affects analysis. *Survey Research Methods*, *16*(2), 227–240.

Sosdian, C.P., & Sharp, L.M. (1980). Nonresponse in mail surveys: access failure or respondent resistance. *The Public Opinion Quarterly*, *44*(3), 396–402.

Stanley, T.D., Carter, E.C., & Doucouliagos, H. (2018). What meta-analyses reveal about the replicability of psychological research. *Psychological Bulletin*, *144*(12), 1325–1346.

Stricker, J., & Günther, A. (2019). Scientific misconduct in psychology. *Zeitschrift Für Psychologie*, *227*(1), 53–63.

Szymanski, D.M., & Henard, D.H. (2001). Customer satisfaction: a meta-analysis of the empirical evidence. *Journal of the Academy of Marketing Science*, *29*(1), 16–35.

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education*, *2*, 53–55.

Towse, J.N., Ellis, D.A., & Towse, A.S. (2021). Opening Pandora's Box: peeking inside Psychology's data sharing practices, and seven recommendations for change. *Behavior Research Methods*, *53*(4), 1455–1468.

Trafimow, D. (2017). Why it is problematic to calculate probabilities of findings given range null hypotheses. *Open Journal of Statistics*, *7*(3), 483–499.

Trafimow, D. (2018). Confidence intervals, precision and confounding. *New Ideas in Psychology*, *50*, 48–53.

Valentine, K.D., Buchanan, E.M., Scofield, J.E., & Beauchamp, M.T. (2019). Beyond p values: utilizing multiple methods to evaluate evidence. *Behaviormetrika*, *46*, 121–144.

Wasserstein, R.L., & Lazar, N.A. (2016). The ASA statement on p-values: context, process, and purpose. *The American Statistician*, *70*(2), 129–133.

Wetzel, M., & Hünteler, B. (2022). The blind spot: studying the association between survey nonresponse and adherence to COVID-19 governmental regulations in a population-based German web-survey. *Survey Research Methods*, *16*(3), 267–281.

Woolston, C. (2015). Psychology journal bans P values. *Nature News*, *519*(7541), 9–9.