Do Employers "Walk the Talk" After All? An Illustration of Methods for Assessing Signals in Underpowered Designs

Jonathan R. Brauer *

Indiana University

Jacob C. Day

University of North Carolina Wilmington

Brittany M. Hammond

Centerstone Research Institute

Direct all correspondence to:

Jonathan R. Brauer, Department of Criminal Justice, Indiana University Bloomington, Sycamore Hall 302, 1033 E. 3rd St., Bloomington, IN 47405-7005, USA.
Email: jrbrauer@indiana.edu

This is an Accepted Manuscript of an article to be published by Sage in a forthcoming issue of Sociological Methods & Research. To cite this article before it is published:

Brauer, Jonathan R., Jacob C. Day, and Brittany M. Hammond. (In press). "Do Employers "Walk the Talk" After All? An Illustration of Methods for Assessing Signals in Underpowered Designs." *Sociological Methods & Research*.

Abstract:

This paper presents two alternative methods to null hypothesis significance testing (NHST) for improving inferences from underpowered research designs. *Post hoc design analysis* (PHDA) assesses whether a NHST analysis generating null findings might otherwise have had sufficient power to detect effects of plausible magnitudes. *Bayesian analysis with default priors* offers advantages over NHST for assessing null findings and detecting signals in underpowered data. Both methods are illustrated by application to Pager and Quillian's (2005) influential study on attitude-behavior correspondence. PHDA results suggest the original study lacked sufficient power to detect strong associations between employers' attitudes and behaviors. Bayesian analysis confirms strong attitude-behavior associations cannot be ruled out given the data. Together, these results question a frequently cited conclusion about attitude-behavior incongruence in survey vignettes. Overall, the examples illustrate how these analytical tools can be useful for describing uncertainty surrounding estimates and for improving substantive and theoretical debates across sociology.

Acknowledgements:

The authors would like to thank Christopher Winship, Jukka Savolainen, Christine Mair, Charles Tittle, Matt VanEseltine, and several anonymous reviewers for comments on previous versions of this manuscript. Also, we learned about the passing of Devah Pager while this paper was under review. Her research was ahead of the curve in its thoroughness, transparency, and replicability, and she has inspired us to strive to live up to her scholarly example.

Author Biographies:

Jonathan R. Brauer is an assistant professor of criminal justice at Indiana University Bloomington. His research tests theories of criminal behavior, explores causes of conflict and discrimination, and examines consequences of coercive and supportive social environments. His research is published in interdisciplinary journals, including *Criminology*, *Journal of Research on Adolescence*, *Justice Quarterly*, *Social Forces*, and *Sociological Quarterly*.

Jacob C. Day, PhD, is an assistant professor in the Department of Sociology and Criminology at the University of North Carolina Wilmington. His research centers on social networks, segregation, and race inequality in labor markets as well as criminal and juvenile justice policy and criminological theory. His work appears in *Work and Occupations, Research in Social Stratification and Mobility,* and *Journal of Research in Crime and Delinquency*.

Brittany M. Hammond received her M.A. from the Department of Sociology and Criminology at the University of North Carolina Wilmington in 2018. She is an Evaluation Associate for Centerstone Research Institute in Bloomington, Indiana. Her interests include racial inequality/stratification and drug policy.

Do Employers "Walk the Talk" After All? An Illustration of Methods for Assessing Signals in Underpowered Designs

INTRODUCTION

Calls for increasing attention to reproduction and replication of scientific knowledge have been present in the social sciences for decades (cf. Freese 2007a, 2007b; Freese and Peterson 2017; King 1995). However, concerns about a "reproducibility crisis" have recently peaked, triggered by numerous failed attempts to replicate high-profile studies across several scientific disciplines (cf. Camerer et al. 2016; Errington et al. 2014; Freese and Peterson 2017; Jasny et al. 2011; Open Science Collaboration 2015). This reproducibility crisis has sparked renewed attention to normative scientific procedures and widespread reporting and publishing biases that may detrimentally affect scientific research (Ioannidis 2005; Simmons, Nelson, & Simonsohn 2011; Gelman and Loken 2015). In response, scholars have called for modifying normative practices and scientific standards to increase the verifiability, robustness, and repeatability of scientific findings (e.g., Freese and Peterson 2017; see also https://www.projecttier.org).

Amidst reproducibility concerns, null hypothesis significance testing (NHST) in quantitative research has been a frequent target of criticism, particularly when such analyses lack sufficient power to detect meaningful effects (Cumming 2014; Fraley and Vazire 2014; Maxwell, Lau, and Howard 2015; Baker 2016). Generally, an NHST analysis involves estimating a focal effect or population difference from a set of observations then calculating a *p*-value, which represents the probability of observing a point estimate at least as extreme as that observed over the long run with repeated random draws of comparable samples from the same population and under the assumption that the null hypothesis (e.g., of no effect or no difference) is true. NHST methods and results are not *ipso facto* problematic; rather, they generate useful

information when applied and interpreted appropriately (Murtaugh 2014; Cumming 2014). However, *p*-values and other NHST results are commonly misinterpreted (cf. Greenland et al. 2016; Wasserstein and Lazar 2016).

Moreover, researchers often apply NHST methods in underpowered research designs (Cohen 1992; Gill 1999; Cumming 2014). Such practices, combined with reporting and publishing biases, may contribute to an overabundance of false positive findings in published research (Ioannidis 2005; Fanelli 2012). Consequently, researchers and editors increasingly are encouraged to report and publish null findings as a remedy to this "file drawer" problem (cf. Franco, Malhotra, and Simonovits 2014; Goodchild van Hilten 2015; Grimes, Bauch, and Ioannidis 2018). Yet, null findings from NHST studies often are misinterpreted as *evidence in favor of the null hypothesis* (Gill 1999; Greenland 2011). Moreover, misinterpretations of null findings are especially problematic in underpowered designs, where such findings both are more frequently expected to occur and more likely to reflect false negatives or Type II errors.

In response, we showcase two methods that can improve conclusions drawn from NHST methods, particularly when null findings are generated in a potentially underpowered design. Though perhaps unfamiliar, these methods should be accessible to most sociologists trained in standard NHST statistical approaches. First, we present a simple *post hoc design analysis* (Gelman and Carlin 2014) for assessing whether conclusions inferred from null findings generated using NHST are likely to be biased by low statistical power. Second, we present *Bayesian analysis with default priors* (Bååth 2014; Gelman et al. 2014, p.51-55) as an important tool for deriving signals from small samples, rare event data, or other underpowered designs.

We illustrate these methods by applying them to published data from Pager and Quillian's (2005) influential sociological study examining the (non)overlap between employers' attitudes

and behaviors. Specifically, the original study reports a null association between employers' decisions to hire applicants with criminal records in a hypothetical survey vignette and their actual callback behaviors in an experimental audit. This null finding is frequently cited as strong evidence of the invalidity of survey methods for assessing real-world behaviors (e.g., Jerolmack and Khan 2014).

The remainder of this paper is organized around providing practical answers to the following methodological and substantive research questions:

1) *How can we know if a NHST analysis generating null findings had sufficient power to detect an effect of a reasonable magnitude?*
2) *How can we detect a meaningful signal from underpowered data?*
3) *How likely is it that Pager and Quillian's (2005) null findings indicate a lack of association between "what employers say" and "what they do?"*

POST HOC DESIGN ANALYSIS

The first question motivating this research is: *How we can know if a NHST analysis generating null findings had sufficient power to detect an effect of a reasonable magnitude?* Below, we briefly describe statistical power and some problems stemming from underpowered designs, then describe *post hoc design analysis* as a useful tool for retrospectively calculating meaningful estimates of the power of a statistical test.

Statistical Power and Inference Errors

Statistical power in NHST is the probability that a given test will correctly reject the null hypothesis. It is determined by a study's (sub)sample size, the estimated effect size, the alpha level, and measurement variance. Generally, studies examining large expected effects with low

variance in large samples have a high probability of correctly rejecting the null hypothesis at a given alpha level.

In an underpowered study, the probability of accurately inferring the existence of a true nonnull effect is low. As a common rule of thumb (see Cohen 1992), a test with power less than .80 is considered underpowered and, hence, overly prone to such Type II errors. Somewhat paradoxically, low statistical power also increases the chances of *overestimating* effect sizes for true nonnull effects. This is because only exaggerated estimates can pass an alpha level threshold in a substantially underpowered study, and "statistically significant" nonnull estimates are more likely to be reported and published than "statistically nonsignificant" estimates (cf. "winner's curse" discussions in Button et al. 2013; Vasishth and Gelman 2017).

Put simply, underpowered studies are characterized by a low probability of *accurately detecting a signal* indicating the existence of a true nonnull effect and a low probability of *detecting an accurate signal* of a true nonnull effect. Certain research designs are notoriously underpowered, such as those investigating small effects or relying on small samples (cf. Bradburn et al. 2007; Oakes 2017; Turner, Bird, and Higgins 2013). However, even studies investigating large expected effects in large samples can be substantially underpowered.

For example, consider the power of a test of group differences in proportions (which we apply later to Pager and Quillian's [2005] data on differences in employer callback rates). Given the same expected effect size, fixed power and alpha levels, and equal sample sizes for both groups, the power of this test increases with the rarity of the outcome (i.e., because variance necessarily decreases as the baseline event rate diverges from $p = .50$). However, observational data often contain imbalanced groups, and power is limited by the smallest subgroup frequency. Moreover, when also examining rare events across imbalanced groups, a test may be comparing

cells containing zero or near-zero frequencies. Hence, NHST analyses may suffer from issues such as biased estimation and lack of statistical power when data contain only a small number of cases on the rarer of two outcomes (King and Zeng 2001; Ma, Chu, and Mazumdar 2016; Vuolo, Uggen, and Lageson 2016), even if the total sample size and expected proportion difference across groups both are relatively large.

Post hoc Power Calculations

Given the high likelihood of inference errors in underpowered research, it is crucial to determine whether a NHST analysis has sufficient power to detect an effect of a plausible magnitude. For this reason, an *a priori* power analysis is routinely recommended before collecting data (e.g., Lakens and Evers 2014). Yet, much quantitative research in the social sciences involves analyses of secondary data, and useful insights are often gleaned from exploratory analyses that were not planned in advance and could not have been subjected to an *a priori* design analysis (Gelman and Loken 2014, p.464-5). Therefore, meaningful *post hoc* power estimates would be useful for evaluating many of the NHST designs and results in our field.

Unfortunately, the most common method for retrospectively estimating statistical power is to use a reported or published effect size estimate from a specific test to calculate the power of that same test. However, such so-called "observed power" or "post hoc power" estimates are fundamentally uninformative (i.e., redundant with the observed effect's *p*-value) and, worse, are frequently misinterpreted (Hoenig and Heisey 2001; O'Keefe 2007).

In contrast, Gelman and Carlin's (2014) post hoc design analysis (PHDA) represents a significant departure from retrospective "observed power" calculations. Centrally, their PHDA calls for the thoughtful consideration of "plausible" effect sizes for a test that are based on external knowledge of substantive relationships (see 2014, p.647-8). Upon positing these

plausible effect sizes, their PHDA method involves calculation of the probability of making errors when drawing inferences from NHST results.

It is important to note that Gelman and Carlin's (2014) PHDA aims to shift focus *away* from power calculations (and statistical significance) and toward effect size estimation and precision by encouraging calculation of the probability of sign (Type S) and magnitude (Type M) errors.[i] With that said, NHST methods are still paradigmatic in the social sciences, statistical power remains a well-known (if often misunderstood) and important concept for NHST analysis, and concerns about potentially underpowered NHST analyses are central to the "reproducibility crisis." Thus, in our view, attempts to retrospectively generate *meaningful* estimates of the statistical power of a NHST analysis are defensible as well as informative in certain situations. We argue the process of generating power estimates can be useful for assessing published conclusions from null NHST findings, and it is most valuable if paired with follow-up Bayesian analyses aimed at estimating credible effect sizes and relative probabilities for null and alternative hypotheses.

Thus, we adopt Gelman and Carlin's (2014, p.647-8) practice of identifying plausible effect sizes, or what we refer to as *counterfactual expected effects*, in post hoc design analysis. The identification of counterfactual expected effects is conceptually akin to the processes involved in proper *a priori* design analysis and the identification of Bayesian informative priors. Moreover, it is this central element of Gelman and Carlin's PHDA that permits calculation of meaningful retrospective power estimates.

Given their direct relationship to *p*-values, power estimates calculated using counterfactual expected effects can serve as an intuitive indicator for whether a NHST study generating null findings *otherwise might have been able to detect an effect of a reasonable magnitude*. Our

application to Pager and Quillian's (2005) data illustrates the utility of this power-centered approach to PHDA in the context of null NHST findings. Here, we recommend the following procedures for conducting a post hoc assessment of the power of a NHST study:

1) Identify the design elements of a given study that are relevant to power calculations, including sample size, variable distributions or measurement variance, and alpha levels.
2) Given the study design features listed in #1 (e.g., sample size and variable distributions), hypothesize a *range* of plausible effect size estimates, or *counterfactual expected effects* for the focal association or population difference.
    a. These counterfactual effects should be independent of the study's observed estimates of effects or population differences. That is, they should be determined using theory and logic in combination with any available *external* information about the focal effects or population differences (cf. Gelman and Carlin 2014, p.647-8; Lakens and Evers 2014).
    b. The focal "observed" effect size or population difference and associated *p*-value may be noted for later comparisons, but do *not* use these observed values in post hoc power calculation.
3) Calculate a range of meaningful post hoc power estimates by entering the design elements identified in #1 and the counterfactual expected effects identified in #2 into a standard statistical calculator or software program designed to calculate power.

Counterfactual Effects: From Testing to (Bayesian) Estimation

Following the procedures above, a PHDA can indicate whether a given NHST study that generated null findings might have been sufficiently powered to detect a range of plausible effects of reasonable magnitudes, thereby answering our first question. Like an *a priori* power

analysis, if the post hoc power estimates from a PHDA are below .80 (Cohen 1992) or an alternative error threshold (e.g., .90; Lakens and Evers 2014), then the test in question might be considered underpowered. A smaller power estimate essentially indicates that, given the study's design features and plausible counterfactual effect sizes, a comparable NHST analysis would have a low probability of correctly rejecting the null (i.e., of detecting a true nonnull effect). In these cases, one should take extra caution in making inferences or drawing conclusions from the statistical test. As explained earlier, underpowered tests too often result in false negative inferences when failing to reject the null (Type II errors; Cohen 1992) *and*, when rejecting the null, too often result in false positive inferences (Type 1 errors; Colquhoun 2014) or in overestimation of true nonnull effects (Type M errors; Gelman and Carlin 2014).

In addition to allowing retrospective calculation of meaningful statistical power estimates, the identification of counterfactual expected effects in PHDA might encourage a focal shift in quantitative sociology from hypothesis testing to effect estimation with quantified uncertainty (cf. Hoenig and Heisey 2001; Cumming 2014; Gelman and Carlin 2014; Kruschke and Liddell 2018a). Moreover, the thoughtful consideration of counterfactual expected effects might serve as a bridge to help scholars trained in NHST methods grasp Bayesian methods that rely on informative priors, since these methods involve a conceptually similar process of identifying plausible or likely effect size estimates (Gelman et al. 2013). As we will demonstrate, these counterfactual effects can also be used as simple heuristic thresholds that enrich interpretations in a simple follow-up Bayesian analysis specifying a default prior distribution.

BAYESIAN ANALYSIS

Recall, the second research question: *How might we detect a meaningful signal from underpowered data?* Bayesian analysis is well-suited for this question, as it provides quantifiable

estimates of both the magnitude of an effect and the degree of precision or uncertainty surrounding an effect estimate. Underpowered designs are characterized by imprecise estimation of effects or population differences; a Bayesian analysis provides a quantified summary in intuitive probability terms of the degree of precision or confidence associated with a range of credible estimates. We begin with a brief description of Bayesian analysis, then describe how a Bayesian approach to data analysis is particularly useful when NHST methods generate null findings, as is often the case in underpowered studies.

Much Ado about Priors

A defining feature of the Bayesian approach to data analysis is the specification of a prior distribution for model parameters. In Bayesian analysis, a posterior probability density is estimated as a function of both the data-driven likelihood estimates and a prior distribution for the model parameters. Specification of a prior distribution allows researchers to intentionally build in substantive knowledge about (a range of values for) a focal effect or population difference (i.e., "informative" priors). Subsequently, the analysis permits assessment of the degree to which model parameters estimated from new observations update – that is, modify or confirm – prior knowledge as represented in the prior distribution. Studies with very large samples and low variation would typically result in (posterior density) estimates that primarily reflect the "signal" in the data (i.e. the likelihood). Conversely, Bayesian estimates (from the posterior distribution) can minimize inference errors (i.e., by relying more heavily on an informative prior distribution) when the signal-to-noise ratio is low, such as when the data (likelihood) generate highly imprecise estimates from small samples with high variability.

Not all applications of Bayesian analysis involve attempts to specify the subjective "state of knowledge" in an informative prior distribution (Gelman et al. 2013, p.34). Rather, many

Bayesian applications rely on so-called "default" prior distributions, which are otherwise known as "noninformative," "reference," "objective," "weakly informative," or "minimalist" priors (cf. Berger 2006; Ghosh 2011; Gelman et al. 2013, pp.51-55; Gelman et al. 2017). Default priors are typically specified to ensure that posterior probability densities are driven primarily by the observed data. An example is the uniform or flat prior distribution for the binomial parameter, which assumes that all plausible values of *y* (e.g., observable "success" rates) are *a priori* equally probable (see Gelman et al. 2013, p.29-34).

Here, we recommend Bayesian analysis with default priors for a few reasons. First, because estimates (the posterior) are dominated by the data (likelihood), this approach typically generates comparable results to NHST methods that are more familiar to sociologists.[ii] Specifically, with large samples, a NHST point estimate usually approximates the median of the posterior density under a default prior specification, and NHST 95% confidence intervals similarly approximate Bayesian 95% posterior density credible intervals (see Schoot et al. 2014). However, with small samples, Bayesian analysis can improve NHST inferences – even with weakly-informative prior distributions – by generating stabilized estimates that are more robust, make more sense, or perform better predictively in certain situations (Gelman et al. 2008).

Relatedly, since a Bayesian analysis with default priors essentially "tips the scales" in favor of the data and generates results that are comparable to NHST results, this approach may be useful for assessing the degree of uncertainty surrounding published NHST estimates. Subsequently, scholars can assess whether such data-driven findings make sense in view of prior knowledge, either by specifying informative priors in a follow-up analysis or, as we will illustrate, by simply comparing estimates and credible intervals to a range of plausible counterfactual expected effects identified in PHDA.

Finally, the proper translation of substantive knowledge into mathematical prior distributions can be complex and may require additional training or consultation. In contrast, default priors are generally easier to use and interpret. For instance, in our example below, we rely on the *Bayesian First Aid* package in R (Bååth 2014), which is specifically designed to ease the transition from NHST methods to Bayesian alternatives. Currently, the package provides alternatives to six classical *.test* functions in R (e.g., binomial test; t-tests; Pearson correlation; test of proportions; Poisson test).[iii] Moreover, Bayesian regression alternatives with default prior specifications are built into in many popular statistical packages, including R (e.g., *bayesglm*; see Gelman et al. 2008), SAS (e.g., *bayes* statement; see Stokes, Chen, and Gunes 2014), SPSS (*Bayesian Statistics* menu)[iv], Stata (e.g., *bayes* prefix)[v], and Mplus (estimation= *bayes*; see Muthén 2010).

Null Findings: No Effect or Weak Signal?

Though widely applicable to many research questions (see Gelman et al. 2013; Kruschke 2015), Bayesian analysis is especially useful in situations where an NHST analysis generates null findings (Kruschke 2011). To illustrate, consider that a null finding in NHST is fundamentally uninformative about the true nature of an effect or population difference. As Gill (1999:661) explains,

> Failing to reject the null hypothesis essentially provides almost no information about the state of the world. It simply means that given the evidence at hand one cannot make an assertion about some relationship: all you can conclude is that you can't conclude that the null was false. (1999:661).

Despite being wholly uninformative, null findings from NHST are routinely misinterpreted as evidence that the null hypothesis is true. For instance, Gill (1999) observed that

between 38% and 51% of reviewed articles in four leading political science journals reporting null hypothesis significance testing ultimately drew substantive conclusions from a fail to reject decision (see also Cohen 1992; Greenland 2011).

In contrast, compared to NHST, Bayesian analysis provides more information about the strength and precision of a given "signal" in the data. A key benefit of Bayesian analysis in these cases is the generation of relative probabilities for various alternative hypotheses, which permit researchers "to quantify to what extent null results reflect a real absence of effects or a lack of statistical sensitivity" (Vadillo et al. 2016, p.88; see also Kruschke 2011; Dienes 2014).

Put differently, in cases where NHST generates null findings, Bayesian analysis permits researchers to determine if the null hypothesis is "more credible" than an alternative hypothesis, which is something "NHST can never do" (Kruschke and Liddell 2018a, p.196). Kruschke and Liddell note this feature "is highly desirable for theoretical domains in which 'proving' the null is the goal" (p.196), as is arguably the case in Pager and Quillian's (2005) study.

EXAMPLE: "WALKING THE TALK"

Devah Pager's (2003) experimental audit is among the most well-known studies in contemporary sociology. Among other findings, her audit study showed that employers were less likely to call back applicants with a (randomly assigned) criminal record compared to similar applicants without criminal records.

Pager and Quillian's (2005; hereafter P&Q) study paired these employer audit data with a self-report survey of the individuals in charge of hiring for audited employers (Pager 2002). Published analyses of the survey data reported stark disparities between self-reported employer attitudes and actual callback practices in the audit study. When presented with a vignette describing an applicant similar to those who had audited the employers, 62% of employers

surveyed expressed some degree of willingness ("somewhat likely" or "very likely") to hire a drug offender who had recently served a prison sentence. However, in the audit study, only 17% of white and 5% of black testers received callbacks.

Perhaps more surprising was the reported lack of association between employers' reported willingness to hire a drug offender and their actual callback behaviors (Kendall's Tau-b = .012, $p > .05$; see P&Q 2005, p.367). This discrepancy between employers' vignette reports and audit behaviors is the central theme of Pager and Quillian's (2005) influential *American Sociological Review* article reanalyzed here, entitled "Walking the Talk? What Employers Say versus What They Do."

Legacy of "Walking the Talk"

P&Q's article has been widely read and cited, accumulating over 500 citations on Google Scholar to date. More importantly than how frequently the article is cited is *how* or *why* it is cited. Despite the original authors' measured discussions and their cautions against overstating conclusions, this article routinely is cited as evidence of the potential invalidity of survey methods for assessing real-world behaviors. Consider a few recent examples from sources published in 2017 expressing concerns about uncertainty surrounding the validity of surveys:

- "A number of significant concerns regarding vignettes' real-life relevance come out of the work of Pager and Quillian (2005)…" (McDonald 2017, p.5).
- "Scholars have shown that in some cases, respondents may seek to give socially appropriate answers to questions, even if this involves distorting the truth (Pager and Quillian, 2005)." (Occhiuto 2017, p.278)
- "…since Pager (2005) unquestionably demonstrated incongruence between what employers say and what they do." (Reich 2017, p.129)

While citations of this kind are often defensive, reflecting authors' attempts to recognize concerns about the use of survey designs, not all citations fit this description. Perhaps the strongest example is Jerolmack and Khan's (2014) recent article in *Sociological Methods and Research*, entitled "Talk is Cheap: Ethnography and the Attitudinal Fallacy." In this article, the authors draw heavily on P&Q to make a case in favor of adopting ethnographic methods and against using "verbal accounts" like those collected in surveys due to the "fact that what people say is often a poor predictor of what people do (Jerolmack and Khan 2014, p.178)."

Why Revisit "Walking the Talk?"

P&Q's study has three particularly desirable characteristics for illustrating the utility of post-hoc design analysis and Bayesian analysis. First, the original study reports a statistically nonsignificant and near-zero association between attitudes and behaviors. Recall, null findings from NHST are uninformative, and underpowered statistical tests (and other design features) can cause failures to reject the null. Nonetheless, P&Q's null finding is regularly cited as evidence of attitude-behavior incongruency, or as evidence *in favor of the null* (e.g., Jerolmack and Khan 2014). A PHDA can help determine whether the initial study design might have had sufficient power to detect a reasonably strong attitude-behavior association in the first place.

Second, the focal study's sample size (n=156) might seem sufficiently large to suppress obvious concerns about low statistical power. However, the focal study analyzed employer callbacks of applicants with criminal records, which is a relatively rare event. Statistical power is limited by small subgroups and imbalanced cells, and these characteristics are common in observational and rare event data routinely analyzed by sociologists (Bradburn et al. 2007).

Third, the focal study is impressively transparent in reporting data and methods, particularly for research conducted prior to the recent replication crisis and subsequent

transparency and open science movements. As a result, it is possible to reproduce the original analyses and verify published conclusions using a Bayesian approach, as well as assess the extent to which researchers' measurement decisions affect the robustness of reported results using data published in the original article. Thus, to a degree, we can assess how the application of NHST methods and the "garden of forking paths" (Gelman and Loken 2014) might have affected P&Q's estimate of the association between employers' attitudes and behaviors.

DATA

Pager's (2003) original audit experiment involved sending same-race matched pairs of white (n=150 pairs) and black (n=200 pairs) men of similar age and credentials to apply for advertised, entry-level employment job openings in Milwaukee, Wisconsin. In total, 350 employers each were audited by a pair of same-race applicants (n=700 applications), with one member of each pair randomly assigned to a "criminal record" condition.

The follow-up employer survey included a vignette closely approximating the audit conditions. Employers were presented with a hypothetical description of a 23-year old black or white male named Chad (with Chad's race matching the auditors' race) applying for an entry-level opening. The vignette described Chad as having good references and interacting well with people, and it indicated Chad was convicted of a drug felony, served 12 months in prison, was released last month, and is now looking for a job. Employers were then asked how likely they are to hire Chad for an entry-level opening. Four response options ranged from "very likely" to "very unlikely." In all, 177 employers completed the survey for a response rate of 51% (Pager 2002). For additional details about the original employer audit or follow-up survey data, see Pager (2003; 2007) and P&Q (2005).

Pooling data from the audit study and the follow-up survey (valid N=156) permitted P&Q (2005) to examine congruence between employers' vignette-based hypothetical hiring decisions and their actual callback behaviors. Our reanalysis relies on cross-tabular group frequency data published in the results and appendices of P&Q's (2005: Table 2, p.367 and Appendix B, p.377) article. Ideally, this reanalysis would employ the original data. However, while the employer survey data were publicly available, the matching employer audit data were neither publicly available nor available upon request.[vi]

MEASUREMENT VARIATIONS

In P&Q's analysis, employers' audit callbacks were measured dichotomously as "yes" or "no" to indicate whether an audited employer called back the audit-matched applicant with a criminal record. Similarly, P&Q collapsed the *very likely* and *somewhat likely* survey responses and collapsed the *somewhat unlikely* and *very unlikely* responses to create a dichotomous indicator of whether an employer was "willing" or "unwilling" to hire the comparable hypothetical applicant.

Theoretically, P&Q justified collapsing *somewhat likely* and *very likely* responses by claiming this grouping corresponds most closely with the audit's behavioral callback measure. Specifically, they argued that a callback "may in fact represent a very low bar of approval" when employers contemplate hiring an applicant (p. 364). However, callbacks might also represent a relatively high bar of approval, as employers might only call back those applicants whom they are "very likely" to hire. Moreover, when faced with competing qualified candidates with and without criminal records – as was the case for each employer in the matched-pair audit design – all other employers might at least somewhat favor candidates with comparable qualifications yet no criminal records.

Therefore, our reanalysis compares results obtained using the original coding condition with those produced using three alternative dichotomous employer contrasts: (1) *very likely to hire* versus *at most somewhat likely to hire*; (2) *very unlikely to hire* versus *at least somewhat unlikely to hire*; and (3) employers who answered *very likely* versus *very unlikely*. These contrasts permit assessment of the robustness of published conclusions across measurement specifications.

EXAMPLE #1: POST HOC DESIGN ANALYSIS (PHDA)

Methods

A key goal of the post hoc design analysis is to assess whether P&Q's data offer sufficient statistical power to detect a vignette-audit association of a reasonable magnitude. For this analysis, statistical power estimates are calculated at alpha=.05 with the Fisher method using the *power.exact.test* procedure in R (version 3.3.3).[vii] In interpreting results, we rely on the commonly recommended power threshold of .80 (Cohen 1992). Analyses falling short of this threshold may be substantially underpowered and prone to Type II errors.

Identifying Counterfactual Expected Effects

The greatest challenge faced in a post hoc design analysis is determining plausible effect sizes (Gelman and Carlin 2014). In the case of P&Q's study, this involves making *a priori* assumptions about how large a difference in audit callbacks of applicants with criminal records one should have expected to observe between employers who say they are more versus less likely to hire such candidates. In other words, to calculate meaningful estimates of statistical power, we must first identify *counterfactual expected effects*, or a range of estimates of the degree of overlap between vignette and audit data that might have been expected prior to conducting the original study.

As explained previously, counterfactual expected effects must be independent of the observed degree of overlap. Here, we estimate what the joint vignette/audit frequency distributions might have looked like if there were a relationship between what employers say and do, given the marginal probabilities of audit callbacks and of employers' vignette-reported willingness to hire candidates with criminal records. Since counterfactual estimates are inherently disputable (for a thoughtful discussion, see Gelman and Carlin 2014:647-8), we start by identifying the "maximum possible" association that could have been observed, then calculate two different estimates representing a "moderate" and a "strong" expected vignette/audit association.

How Large an Effect Might Have Been Expected?

It is tempting to assume that a high correlation coefficient is expected between vignette and audit results, such as is found in prior research assessing attitude-behavior correspondence (e.g., $r > .50$; cf. Glasman and Albarracin 2006; Vaisey 2014). However, this assumption is misleading when estimating expected associations between binary variables measuring rare events, such as job applicant callbacks.

P&Q's (2005:377) conclusions about group differences in callback rates are based on only *eleven* total observed callbacks of applicants with criminal records across all 156 employers with valid overlapping survey and audit data. Since the vast majority of employers (93%) in both the "more willing" and "less willing" groups do not call back candidates, these shared non-events suppress the maximum size of observed associations. Furthermore, without a "no criminal record" control condition in the survey vignette, P&Q's (2005) analyses are limited to examining callbacks of candidates *with* criminal records – an especially rare event.

As an illustrative example of the effect of imbalanced frequency distributions on the calculation of correlation coefficients, consider the size of the "criminal record effect" on callbacks documented in Pager's (2003) audit study. The marginal probabilities of callbacks for whites and blacks with and without a criminal record (Pager 2003:958) can be transformed into a frequency distribution, from which a correlation coefficient can be calculated. For instance, it appears that 79 of the 350 applicants in the "no criminal record" received callbacks (whites: 150*.34=51; blacks: 200*.14=28), compared to about 36 of the 350 applicants in the "criminal record" condition (whites: 150*.17=25.5; blacks: 200*.05=10). This "criminal record effect" equates to an odds ratio of 2.54, or a Phi (Φ) coefficient of association of -0.17.

This coefficient might be mistakenly interpreted as a small effect, for instance, if one were to rely on Cohen's (1988; 1992) "rules of thumb" for interpreting correlational effect sizes. Such a mistake becomes apparent upon calculating the "maximum possible" Phi coefficient (Φ_{max}) for the observed marginal distribution. This can be done simply by imagining that *all* 36 of the applicants in Pager's criminal record condition who received callbacks from employers instead *did not receive callbacks*. In this counterfactual case, the Phi coefficient representing the correlation between "criminal record" condition and employer callbacks would equal a maximum negative value of -0.35. In other words, this is the lower-bound (i.e, "maximum" negative) value that Phi might have taken given marginal probabilities. Note that this value deviates substantially from -1.0, or the lower standardized threshold for a Pearson's *r* coefficient.

Thus, as illustrated in this example, rare events suppress the minimum and maximum possible values for Phi, making interpretation of effect sizes such as Phi values (e.g., -0.17) problematic using widely known "rules of thumb" (Cohen 1988).[viii] However, dividing observed Phi by maximum Phi (Φ/Φ_{max}) produces a normed coefficient that is conceptually and often

empirically comparable to the absolute value of a Pearson's *r* (but not equivalent; cf. Kaltenhauser and Lee 1976:310; Davenport and El-Sanhurry 1991; Olivier and Bell 2013). Using this normed statistic, the magnitude of the observed "criminal record effect" in Pager's (2003) audit is approximately half the maximum possible given the observed marginal probabilities (Φ/Φ_{max} = -.17/-.35 = .49), which is conceptually akin to Cohen's (1988) threshold for a "large" or strong effect (*r* = .50).

As in the above illustration, we begin our post hoc design analysis by identifying counterfactual "maximum possible" vignette-audit associations ($\Phi/\Phi_{max} = 1$) for each coding condition. Subsequently, based on prior information about meta-analytic average correlations between survey-based and observational measures of behavior (Glasman and Albarracin 2006; Kraus 1995), these "maximum" distributions are modified to produce two counterfactual expected callback distributions approximately corresponding to "moderate" ($\Phi/\Phi_{max} \approx .38$) and "strong" ($\Phi/\Phi_{max} \geq .52$) vignette-audit associations. Estimation of these counterfactual effects relies on knowledge about probabilities of the outcome, or the relative rarity of audit callbacks, as well as information about probabilities of treatment/group allocation, or different levels of vignette-reported willingness to hire applicants with criminal records (For a discussion of these two types of probabilities and an example of their use in a reanalysis, see Olivier and Bell 2013). Probabilities of the outcome are derived from observed callback proportions in Pager's (2003) original audit study; probabilities of group allocation are marginal probabilities derived from survey responses in P&Q's (2005:377) analytic sample.

Assumptions about Expected Callbacks

We calculate expected callback proportions for employers in the "more willing to hire" groups in each coding condition by multiplying the marginal probability for the "more willing"

group by a weighted estimate of the probability of callbacks for that group. First, to calculate the weighted probability of a callback, we assume that employers who report being *very likely* to hire the hypothetical applicant with a criminal record in the survey vignette will, on average, call back candidates with criminal records at the audit-average callback rate for applicants *without* criminal records. That probability, combined for whites and blacks, is .226 ([.34*150 + .14*200]/350 = 79/350 = .226; see estimates from Pager 2003:958). This expected callback probability for employers in the *very likely* group is an "anti-conservative" estimate that should result in "conservative" statistical conclusions favoring the original study. Specifically, it is untenable to assume that employers, when faced with competing candidates, are as likely to call the applicant with a "criminal record" as they are to call the audit-matched candidate without a record. Hence, the plausible expected effect sizes generated by this analysis should be inflated and, thus, should tip the scales in favor of supporting P&Q's conclusion that the vignette/audit overlap is weak or null in these data.

In comparison, we assume that employers who report being *somewhat likely* and *somewhat unlikely* to hire the hypothetical drug offender will call back applicants with criminal records at the audit-average rate that employers called back candidates in the "criminal record" condition. That callback probability, combined for whites and blacks, is .103 ([.17*150 + .05*200]/350 = 36/350 = .103]. Though perhaps more realistic than the *very likely* assumption, this assumption should also be somewhat anti-conservative in that it presumes employers reporting some reservations (*somewhat likely*) and those reporting even greater reservations (*somewhat unlikely*) about hiring a hypothetical candidate with a criminal record both will call back such applicants at the audit-average rate in real-world competitive hiring conditions.

Based on these assumptions, we might have expected approximately five callbacks of ex-convicts from the 22 employers in the *very likely* group (.226*22 = 4.97), eight callbacks from the 74 employers in the *somewhat likely* group (.103*74 = 7.62), and three callbacks from the 28 employers in *somewhat unlikely* group (.103*28 = 2.88). This equates to 16 expected callbacks out of 156 applications in the "criminal record" condition, or an expected callback probability of .103. This counterfactual callback rate (.103) is equivalent to the audit-average callback rate observed in Pager's (2003) study and greater than the observed callback rate in the overlapping vignette-audit sample (11/156 = .071).

Calculating "Maximum Possible" Associations

After fixing the marginal probabilities for vignette responses and setting assumptions about expected audit callbacks among employers in the "more likely" groups, we can calculate the "maximum possible" association (Φ_{max}) for each vignette coding condition. This is achieved by fixing the expected number of audit callbacks by employers in the "less likely" group to zero. For example, using P&Q's (2005) coding, this procedure results in 13 expected callbacks out of 96 employers in the *somewhat/very likely* group (*somewhat likely* callbacks = 8; *very likely* callbacks = 5) versus zero expected callbacks out of 60 employers in the *somewhat/very unlikely* group (see Column 1, Panel 1.A, Table 1 in Results below), for a maximum Phi association (Φ_{max}) equal to .24.

"Strong" and "Moderate" Associations

Counterfactual "strong" and "moderate" frequency distributions were estimated by fixing the marginal probabilities for vignette group membership and callback outcomes at the same values in the "maximum possible" distributions for each coding condition, then adjusting expected callback frequencies to identify distributions where Φ/Φ_{max} is, respectively, equal to or

greater than .52 (“strong”) or approximately equal to .38 (“moderate”). These values were selected based on prior information about comparable meta-analytic average correlations between survey-based and observational measures of behavior (Glassman and Albaraccin 2006; Kraus 1995). This decision is in line with both Cohen and Cohen’s (1983:59-60) and Gelman and Carlin’s (2014:647) recommendations for hypothesizing an expected effect size. Moreover, these normed values have the added advantage of being conceptually comparable to Cohen’s (1988) “rule of thumb” thresholds for medium (r = .30) and large (r = .50) effect sizes in the social sciences.

This procedure generates counterfactual “strong” distributions (Column 2 in Table 1) with Φ/Φ_{max} values ranging from .53 to .70. Examination of the relative odds of expected callbacks for employers in the “more likely” versus “less likely” groups in these “strong” distributions reveals odds ratios ranging from 3.75 to 10.42 (see Table 1). Each of these values exceed Olivier and Bell’s (2013) recommended “strong” threshold (OR = 3.0) for 2x2 tables.

The counterfactual “moderate” distributions (Column 3, Table 1, in Results below) were identified by changing the "strong" distribution by one callback. That is, we moved a single expected callback from "more likely" to "less likely" in each coding condition. This change resulted in Φ/Φ_{max} values ranging from .30 to .40. Compared to "strong" distributions, these "moderate" distributions may be more plausible – yet still anti-conservative or inflated – expected effect size estimates. This is because design features of P&Q's (2005) study, including the dichotomous outcome variable and dissimilar target behaviors (i.e., hiring versus callbacks), tend to suppress overall attitude-behavior correspondence (for detailed discussions, see Kraus 1995; Glasman and Albarracin 2006).

With that said, even these "moderate" associations equate to relatively large odds ratios ranging from 1.91 to 4.37. Hence, our lower-bound expectation sets a relatively high bar for the vignette-audit overlap, as a "moderate" association is defined as the following expectation in this analysis: An employer who reports being "more willing" to hire a hypothetical candidate with a criminal record in a non-competitive vignette is expected to call back such applicants in a real-world competitive hiring condition – in which the employer faced at least one equally competitive audit-matched candidate without a criminal record – at approximately two to four times the rate of their "less willing" counterparts.

Results

Table 1 presents counterfactual frequency distributions representing "maximum possible" (Column 1), "strong" (Column 2), and "moderate" (Column 3) associations between employers' vignette reports and audit callbacks for each coding condition. Specifically, Panel 1A employs P&Q's (2005) coding condition, whereas the remaining panels contrast employers who are *very likely* versus *at most somewhat likely* (1B); *very unlikely* versus *at least somewhat unlikely* (1C); and *very likely* versus *very unlikely* (1D) to hire the hypothetical applicant.

[TABLE 1 HERE]

Expected vignette-audit associations are summarized using raw (Φ) and normed (Φ/Φ_{max}) Phi coefficients of association as well as proportion differences and odds ratios contrasting audit callbacks between "more willing" versus "less willing" groups. The expected *proportion difference* estimates in the "strong" and "moderate" conditions (Columns 2 and 3) will be used later as benchmarks for evaluating the observed effects in our reanalysis.

In the PHDA, expected effect size thresholds and corresponding frequency and proportion distributions permit retroactive calculation of statistical power. Hence, Table 1 also

includes statistical power estimates for each expected effect magnitude (columns) and vignette coding condition (panels). These estimates reveal a few particularly noteworthy findings.

First, only two of the twelve counterfactual distributions show sufficient power to detect an effect of the displayed magnitude. One such sufficiently powered condition involves P&Q's coding condition (.99; Panel 1A, Column 1); the other involves a contrast between employers in the *very likely* and *at most somewhat likely* groups (.98; Panel 1B, Column 1). In both cases, these results suggest that future analyses using the same design would be sufficiently powered to detect the "maximum possible" effect size at greater than .80 probability. Specifically, using a two-tailed test at alpha=.05, an analysis performed with P&Q's coding should have a 99% chance of observing a statistically significant group difference in callbacks of .14, or the maximum possible observable proportion difference given marginal probabilities. In contrast, analyses involving the remaining two coding conditions (Panel 1C and 1D) might be insufficiently powered, according to conventional standards, to detect a statistically significant group of even the "maximum possible" effect size at greater than .80 probability (power = .66 and .78, respectively).[ix]

Moving on to our plausible counterfactual expected effects, the PHDA results suggest comparable analyses would be severely underpowered to detect a "moderate" effect (power ranging from .06 to .28) as well as underpowered to detect a "strong" effect (power ranging from .22 to .66) in all four coding conditions. Hence, future studies using the same design would have substantially less than an 80% chance of detecting a *moderate* or *strong* association between vignette reports and audit behaviors at α=.05, irrespective of measurement. In other words, even if the true vignette-audit associations were strong in magnitude, these low counterfactual power estimates suggest that NHST studies employing the same design would often generate false

negatives (and, in cases where the null is accurately rejected, analyses would likely overestimate the true nonnull associations). Overall, results of this PHDA suggest that P&Q's study likely lacked the necessary statistical power at the outset to detect a reasonable association between what employers say and what they do.

EXAMPLE #2: BAYESIAN ANALYSIS

Methods

After conducting the PHDA, we verify and assess robustness of P&Q's NHST results (2005:367) by assessing whether employers' vignette-stated willingness to hire a hypothetical applicant with a criminal record are statistically independent of employers' callbacks of such applicants in a matched experimental audit design across multiple measurement conditions. We then extend these NHST analyses by estimating plausible parameter values for vignette-audit association using the *Bayesian First Aid* package in R (Bååth 2014).

Specifically, we use the *bayes.prop.test* command in the *Bayesian First Aid* package. This procedure relies upon observed data and a default prior distribution to estimate relative frequencies of success across groups, in this case proportion differences in audit callbacks across employers who report being *more likely* versus *less likely* to hire an applicant with a criminal record. The package specifies a binomial distribution, with the prior distribution of the relative frequency of success (θ) specified as flat or uniform ($\theta \sim$ Beta[1,1]).[x]

A choice of default or "minimalist" priors (see Gelman et al. 2017) should be appropriate for these simple binomial models, where both successes (callbacks) and failures (non-callbacks) are possible outcomes and prior expert knowledge on the distribution of callbacks and the specific overlap between employers' vignette reports and audit behaviors is sparse.[xi] More importantly, the use of informative priors – such as knowledge of meta-analytic average

correlations between attitudes and behaviors (e.g., $r > .50$; cf. Glasman and Albarracin 2006; Vaisey 2014) – would cause those previously documented stronger associations from larger overall samples to dominate the posterior distribution, thus tipping the scales toward rejecting Pager and Quillian's (2005) conclusions.

Given the observed data and default prior, the *bayes.prop.test* command infers a posterior probability density of relative frequencies (proportion callbacks) from which point estimates (e.g., the median of the posterior density) and 95% credible intervals (highest posterior density intervals, or HDIs) are calculated. In addition, the analysis estimates a posterior probability density, a point estimate, and credible intervals for the *difference* in relative group frequencies.

These estimates can then be compared to the counterfactual expected "moderate" and "strong" effects for each coding condition. Specifically, akin to the logic of null hypothesis testing, we can examine whether expected group differences representing "moderate" or "strong" associations fall within the 95% credible intervals and, hence, are considered plausible values given the data and default priors.

Finally, and in important contrast to null hypothesis testing, these Bayesian estimations of posterior probability densities allow for calculation of the probability that the underlying relative frequency of callbacks is greater in one group compared to another. Put simply, this probability estimate tells us whether it is a "good bet" to assume that employers who say they are *more likely* to hire ex-convicts (in the vignette) do indeed call back ex-convicts (in the audit) at a greater rate than do their *less likely* counterparts. Probabilities close to .50 suggest a poor bet; one is as likely to be wrong as right in betting that employers do what they say given these data (and default priors). Probabilities approaching .99 indicate a good bet, suggesting a high probability that

employers who say they are more likely to hire ex-convicts also are more likely to call back applicants with criminal records.

In this analysis, we calculate three different probability values for each coding condition. Specifically, we present probabilities that the difference between the "more" and "less" willing group is: (1) greater than zero; (2) equal to or greater than the moderate counterfactual effect estimate; and (3) equal to or greater than the strong counterfactual effect estimate.

Results

Robustness of NHST Null Findings

Despite concerns about statistical power, for comparison purposes, Table 2 reports results of null hypothesis tests (Column 3) applied to P&Q's observed joint frequency data (Columns 1 and 2). For each coding condition (Panels 2A–2D), these tests assess whether we can reject the null hypothesis of no difference in audit callback rates between employers who report being "more willing" versus "less willing" to hire applicants with criminal records. Column 3 of Panel 2A in Table 2 reproduces P&Q's (2005:367) published null findings using their original coding. The next three panels (2B–2D) replicate this null finding in all three alternative coding conditions. Given the aforementioned lack of statistical power to detect a realistic association in these data, the null hypothesis test results are unsurprising.

[TABLE 2 HERE]

Bayesian Estimated Group Differences

Column 4 of Table 2 also reports estimated callback proportions or relative frequencies (Rel. Freq., or the median of the posterior distribution) and 95% credible intervals for these estimates (95% highest density intervals or HDIs) derived from Bayesian reanalysis of these data. In addition, Column 5 reports estimates of the *difference* in relative group frequencies (i.e.,

Est. Diff.) and 95% HDIs for these estimated group differences. These values reported in Column 5 are also displayed in Figure 1, which shows full posterior probability densities for differences in callback proportions between "more willing" and "less willing" employer groups.

[FIGURE 2 HERE]

Panel 2A in Table 2 shows the estimated proportion difference in callbacks between employers who report being *somewhat or very likely* and those who report being *somewhat or very unlikely* to hire ex-convicts is 0, with a 95% HDI ranging from -.09 to .08. Two things are particularly noteworthy about the findings reported in Panel 2A. First, as expected, these estimates essentially reproduce Pager and Quillian's published null findings, which showed a failure to reject the hypothesis that employers' vignette responses and audit behaviors are statistically independent (*proportion difference* = .01; 95% CI [-.09, .09]; see 2005:376).

Second, the 95% credible interval for this estimated group difference (Table 2, Panel 2A, Column 2) contains both the counterfactual "moderate" (.05) and "strong" (.08) effect sizes. This is also visually apparent in Panel 2A of Figure 1, as the vertical dashed lines representing "moderate" (a) and "strong" (b) proportion differences both fall within the solid horizontal line representing the 95% credible interval. Hence, although these data cannot rule out the possibility of absolute statistical independence between employers' vignette reports and audit callbacks, neither can they rule out the possibility of a *strong* vignette-audit association. Put differently, using conventional confidence thresholds with these data, we cannot reject the possibility of a vignette-audit association equivalent to an odds ratio of nearly four (*OR* = 3.75; see Panel 1A, Column 2, Table 1) or comparable in magnitude to strong meta-analytic average correlations between survey-based and observational measures of behavior (i.e., Φ/Φ_{max} = .60 *vs.* *r* = .52; see Glasman and Albarracin 2006).

In comparison, results of reanalysis using alternative coding decisions for collapsing employer vignette responses into *more likely* and *less likely* employer groups show larger non-zero estimates of the group differences in callback proportions (ranging from .03 to .06; see Column 5 of Panels 2B–2D in Table 2). Similarly, these results show that moderate and strong counterfactual effect sizes also fall well within the 95% credible intervals of the estimated group differences for every coding condition (see Panels 2B–2D in Figure 1).

Estimated Posterior Probabilities

Thus far, reported results should resemble the familiar logic underlying classic null- or point-equivalence hypothesis tests. In contrast, the probabilities reported in Column 6 of Table 2 represent a salient departure from this frequentist paradigm, and one that illustrates a key benefit of adopting even a simple Bayesian modeling approach specifying default priors to conduct routine statistical tests. These estimates provide a direct, if tentative, answer to the central question motivating P&Q's study – whether there is a mismatch between "what employers say versus what they do." Calculation of these probabilities are made possible by the estimation of the posterior density from the (default) prior distribution and the data-driven likelihood.

Column 6 in Panel 2A, which reproduces P&Q's coding, reports a probability estimate of .52. Consistent with their oft-cited conclusions, this estimate suggests that betting on employers doing what they say is associated with odds that are slightly better than a coin flip. That is, there is an estimated 52% chance that employers who say they are *somewhat* or *very likely* to hire a hypothetical applicant with a criminal record in the vignette also call back applicants with criminal records in the audit study more often than employers who report they are *somewhat* or *very unlikely* to do so. This contrast is visually apparent in Panel 2A of Figure 1, where 52% of the area under the curve (blue shaded region) falls to the right of zero.

In contrast, when employing alternative coding decisions for employers' reported willingness to hire applicants with criminal records (see Table 2, Panels 2B–2D, Column 6), the probability that the relative frequency of callbacks is greater among the *more likely* group compared to the *less likely* group is higher, ranging from .76 to .81. Likewise, in Panels 2B–2D of Figure 1, between 76% and 81% of the area under the posterior density curves falls to the right of zero. In other words, when employer "willingness to hire" is coded differently than in the original study, the probability is greater than 75% that employers who say they are more willing to hire applicants with criminal records indeed do call back such candidates at a greater rate than their less willing counterparts.

Finally, for each coding condition, similar posterior probabilities are reported for the counterfactual expected effect thresholds to the right of Figure 1. These values indicate how likely it is, in probability terms, that there is a "moderate" or "strong" association between employer vignette reports and audit behaviors, given these data and no prior substantive knowledge (i.e., flat priors). Probabilities of a "moderate" association vary between .13 for P&Q's coding condition (Panel 2A) and .45 for the *very likely*/*very unlikely* contrast (Panel 2D). Probabilities of a "strong" association vary between .03 (Panel 2A: Pager and Quillian's coding condition) and .16 (Panel 2B: *very likely*/*at most somewhat likely* condition). While these values offer additional information about the observed vignette-audit associations, we offer them primarily for transparency purposes. We caution against over-interpreting these values, given the anti-conservative or inflated nature of the counterfactual estimates and the unreliability of observed estimates produced by an underpowered study.

DISCUSSION

This paper introduces two methods for assessing and addressing issues emerging from the standard application of NHST analyses in potentially underpowered research designs. First, *post hoc design analysis* (PHDA) is presented as a means for retrospectively assessing whether a NHST analysis that generates null findings otherwise might have had sufficient statistical power to detect a range of plausible and potentially true nonnull effects. Second, *Bayesian analysis with default priors* is described as more informative than NHST for detecting signals in underpowered data. We illustrate the utility of these methods by applying them to Pager and Quillian's (2005; P&Q) influential study, which documented a mismatch between what employers say and what they do when considering hiring applicants with criminal records. Thus, in addition to showcasing PHDA and Bayesian analysis, a primary goal was to assess the validity and robustness of the frequently cited conclusion that what people say they will do in a survey is incongruent with what they do in a real-world audit.

PHDA, Statistical Power, and Null Findings

First, we conducted a post hoc design analysis to identify plausible counterfactual effect sizes and estimate statistical power for the original study. Results from the PHDA (Table 1) suggest that P&Q's initial study likely lacked sufficient statistical power to detect moderate or even strong associations between employers' attitudes and behaviors. This lack of power reflects both a modest overlapping audit and survey sample size (N=156) and the relative rarity of successful callbacks ($N_{callbacks}$=11). In short, the "signal to noise" ratio is weak, making sign and magnitude errors likely in estimating the relationship between what employers say and what they do. (For a detailed discussion of power problems in audit designs, see Vuolo et al. 2016).

Second, we reanalyzed P&Q's (2005) cross-tabular frequency data using NHST methods across four different coding conditions. These robustness checks confirmed P&Q's key finding – in all four measurement conditions, NHST tests failed to reject the null hypothesis of no difference in audit callbacks across employer groups.

Although our NHST reanalysis reproduces P&Q's null result across four measurement conditions, we cannot echo their substantive interpretations of these results. For instance, P&Q (2005:373-4) conclude that the "low correlation between expressed and observed hiring outcomes presents an epistemological worry" and that "these findings suggest that sociologists may need to reevaluate what is learned from studies that use vignettes of hypothetical situations." Yet, strictly speaking, these findings only indicate that we are *unable to reject* the null hypothesis of no relationship between employers' vignette responses and their audit behaviors – they do not provide *evidence in favor of* the null hypothesis.

Rather, failures to reject the null routinely occur even when a true nonnull association exists; such false negatives are especially likely in underpowered designs. After all, a typical social scientific research study that specifies an α=.05 error threshold and power=.80 tolerates a 5% false positive error rate (α-errors) and a 20% false negative error rate (β-errors; power=1–β). Thus, the typical design is *four* times as likely to incorrectly fail to reject the null when a true nonnull effect exists as it is to falsely reject a true null hypothesis. False negatives are even more likely to occur when a study is underpowered. For example, if α=.05 and true power = .50 then false negatives are *ten* time more likely than false positives (cf. "threshold asymmetry;" Burt et al. 2017, p.474). Likewise, our PHDA results suggest P&Q's research design is particularly susceptible to false negatives – it lacks sufficient power to reliably detect a strong, or in some conditions even the maximum possible, association between vignette reports and audit behaviors.

Our PHDA application has broad implications for recent debates about the "reproducibility crisis" in science. One concern emerging from these debates is the scientific community's apparent aversion to publishing null findings. Ferguson and Heene (2012, p.558) referred to this aversion as "…arguably one of the most pernicious and unscientific aspects of modern social science (Ferguson and Heene, 2012:558)." In response, researchers and editors increasingly are encouraged to publish null findings to reduce reporting biases, improve accuracy in meta-analytic results, and ultimately encourage theoretical falsification (Ferguson and Heene, 2012).

From this perspective, P&Q's (2005) study offers a prominent and widely cited counter-example of a null finding published in a high-impact journal (see also Greenland 2011; Vadillo et al. 2016). If researchers and editors begin heeding calls to publish null findings in response to concerns about scientific reproducibility, then studies like P&Q's might become increasingly common across the social sciences. However, if false positives pose a problem for science (Ioannidis 2005; Simmons et al. 2011), then false negatives pose an equal or greater threat to scientific inquiry. As Fiedler and colleagues note (2012, p.663), "[e]very α error [false positive] on a focal hypothesis entails β errors [false negatives] on alternative hypotheses, just as for every falsely convicted person one (or more) true criminals go free." Also, false negatives are less likely than false positives to stimulate subsequent research – and thus less likely to be corrected by failed replication attempts. By lingering around longer, false negatives can perpetually contaminate scientific reasoning. Moreover, scientific advances often emerge from researchers overcoming false negatives, such as through precise specification and crucial tests of innovative alternative hypotheses. Conversely, though comparatively more common, theoretical innovations "hardly ever arise from abandoning false positives" (Fiedler, Kutzner, and Krueger 2012, p.666).

Thus, though we support reporting and publishing null findings, our discussions and empirical example should highlight the importance of critically assessing and cautiously interpreting such findings; PHDA can help achieve this goal. With that said, Gelman and Carlin (2014, p.643) caution against improperly using retrospective power analysis "as an alibi to explain away nonsignificant findings" (cf. Hoenig and Heisey 2001; Greenland 2012). Instead, these authors suggest their version of PHDA is particularly useful for assessing whether strong (statistically significant) nonnull effect estimates might be potentially biased due to an underpowered design (p.642). These cautions notwithstanding, our application to P&Q's study illustrates how a power-focused PHDA using counterfactual expected effects can be used to assess the appropriateness of conclusions from published null findings and determine whether a test might have been underpowered in the first place.

Weak Signals and Bayesian "Best" Bets

Considering the inherently uninformative nature of the null findings from NHST analysis, we presented Bayesian analysis with default priors as a useful method for detecting signals in underpowered designs. In our application to P&Q's data, we show how a Bayesian approach might uncover additional information about the relationship between what employers say versus what they do, even though the underlying signal in the likelihood is weak. Moreover, the Bayesian reanalysis also assessed robustness of conclusions across four measurement conditions.

Overall, results of the Bayesian reanalysis suggest that substantive conclusions about whether employers "walk the talk" may depend on how survey vignette responses are coded. Using P&Q's original coding (i.e., collapsing the *very likely* and *somewhat likely* survey responses), our analysis reproduces their primary conclusions and calls into question whether employers act in accordance with their vignette-reported intentions. Even here, though, a caveat

is necessary. In the original coding condition, a strong association remains plausible by conventional statistical standards, if perhaps unlikely given Bayesian posterior probabilities.

In contrast, if employer willingness is measured differently, then the findings present a challenge to the central conclusion emerging from P&Q's study. Results using alternative coding decisions indicate there is between a 76% to 81% chance that employers who say they are "more willing" to hire indeed call back candidates with criminal records at a greater rate than their less willing counterparts. Hence, even in an underpowered analysis, in which conservative assumptions favoring the original study's conclusions are applied or prior knowledge about attitude-behavior correspondence is ignored altogether, results in three of four conditions suggest it is likely that employers do "walk the talk," though to what degree remains unknown.

The Bayesian models in this study specify default or minimalist priors, which place a greater burden than informative priors on the observed data to detect "signal" from noise. As explained previously, using informative priors such as average correlations in meta-analyses of attitude-behavior correspondence would have resulted in the typically large average prior correlations (e.g., $r = .52$; Glasman and Albarracin 2006) to dominate the posterior distribution, thus tipping the scales toward concluding that employers' attitudes and actions substantially overlap. Yet, this study raises serious questions about the amount of detectable signal present in these data, given the relatively small sample containing very few "success" events (i.e., only 11 total callbacks of applicants with criminal records were observed out of 156 observations).

Given these concerns, we caution against making strong claims from any analyses of these data. Our "best" bet is that employers who say they are very likely to hire applicants with criminal records probably do call back such applicants more frequently than their counterparts. Since P&Q's (2005:377) published data alone do not provide strong grounds for making a

"good" or "safe" bet of any kind, our bet is informed by our Bayesian reanalysis of these data and our prior knowledge of research on attitude-behavior correspondence (cf. Schuman and Johnson 1976; Kim and Hunter 1993; Glasman and Albarracin 2006; Vaisey 2014). Again, had we relied upon this existing research to specify an informative prior distribution or set of distributions for our Bayesian analysis, the strong priors would have dominated the weak signal in the likelihood, and this is the conclusion we would have reached.

On the broader issues of reproducibility and validity of scientific inquiry, we note that even our simple Bayesian application to underpowered data generated more information than do many complex NHST applications to large datasets. This is because common NHST applications in sociology merely involve a simple dichotomous decision: to reject or not to reject. Such point-null tests provide very little information about a single alternative hypothesis and a (generally untenable) null hypothesis: a difference or effect estimate is either "significant" (and worth reporting) if the *p*-value is less than a specified threshold (e.g., α=.05), or it is not. This "yes/no" logic both parallels and reinforces simplistic "theoretical" debates in sociological literatures, which are often framed around whether two groups differ or whether a construct, variable, or process has an "effect" on an outcome of interest. Such simplistic debates too easily devolve into ideological shouting matches that fail to meaningfully advance substantive knowledge or theoretical precision about important social scientific questions.

In contrast, our Bayesian analysis easily generated an entire posterior probability density of effect estimates that can be used to compare the tenability of various alternative hypotheses. In our example, we showed how Bayesian analysis permits estimation of the probability, given priors and the likelihood, that the association between employers' attitudes and behaviors is greater than zero; using the same analysis, we were also able to estimate the probability that the

association is at least moderate or strong in size. While data limitations preclude precise estimation in this case, as estimation becomes more precise and credible intervals shrink in sufficiently powered analyses, even more informative comparisons are possible. For instance, one can define a "region of practical equivalence" or ROPE (Kruschke 2011, p.302) around zero by identifying a range of effect sizes that are deemed small enough to be substantively equivalent to zero. Then, if the credible interval around an effect estimate falls entirely outside the ROPE, one can plausibly reject the hypothesis that the effect is practically negligible in magnitude. Alternatively, if the credible interval is entirely inside the ROPE, one can essentially *confirm the null* – technically an impossibility in NHST – by concluding that the effect in question is likely zero or practically negligible (see Kruschke 2011).

Overall, Bayesian analysis is specifically designed to quantify the precision (or credibility) around our estimates. Reporting and interpreting confidence intervals from a classical frequentist analysis – as P&Q (2005) do in parts of their study – similarly shifts emphasis from significant/nonsignificant hypothesis test decisions to estimation and precision (Cumming 2014). While such practices arguably represent improvements over common NHST procedures, results can also diverge sharply from Bayesian posterior distributions and generate misleading interpretations, particularly with small samples (cf. Gelman et al. 2013: p.91-95; Kruschke and Liddell 2018a).

Moreover, when combined with informative priors or counterfactual expected effects as presented here, Bayesian analysis also requires researchers to be clear and precise about their subjective evaluation of the state of evidence regarding a given phenomenon. Hence, the methods advocated for and demonstrated here can help "nudge" sociologists to more fruitfully advance substantive and theoretical debates in the face of the uncertainty and imprecision that

characterizes much of our results. First, these methods might help avoid misinterpretations and overreactions to uncertainty stemming from underpowered studies, such as those frequently exemplified in pessimistic interpretations of P&Q's study (e.g., Jerolmack and Kahn 2014). Second, these methods might generally improve our ability to summarize the uncertainty inherent to foregoing the laboratory and relying on noisy data from people living in an uncertain, changing, and imprecise world. Third, such methods might help us avoid "arrogance" and other problems caused by the failure to recognize or be transparent about what we *do* and *do not* know when engaged in theoretical and policy-oriented debates (Tittle 2004). Finally, a focal shift away from NHST and toward positing effect sizes, quantifying uncertainty, and contrasting alternative hypotheses might encourage *theoretical* movements toward precision and "strong inference." Hence, in advocating for these alternative methods, we echo Fiedler and colleagues' (2012, p.667) conclusions:

> "The growth of science depends not so much on technical procedures of significance testing, but on clearly articulated theories and upfront debates leading to crucial tests of alternative hypotheses. Real progress can only be attained when clearly spelled-out theories enable and force researchers to predict what empirical results a theory excludes and what evidence might falsify a given theory or, preferably, allow for clear-cut decisions between two or more competing theories."

CONCLUSION

The present study introduces post hoc design analysis (PHDA) and Bayesian methods as tools that can generate valuable information beyond the results generated by standard NHST statistical approaches. We illustrate how these methods are especially useful for improving statistical inferences and minimizing errors in conclusions derived from underpowered designs.

In an example, we show how pessimistic conclusions stemming from P&Q's (2005) published null finding about the potential invalidity of survey vignettes – and of surveys more generally – are premature. Finally, we argue that adoption of these analytical tools might help push the field of sociology towards more sophisticated theoretical and substantive debates, which we view as necessary for meaningful scientific progress given the complexity of topics in our field and the impact that many sociologists want to have in the social world.

REFERENCES

Bååth, Rasmus. 2014. “Bayesian First Aid: A Package that Implements Bayesian Alternatives to the Classical *.test Functions in R.” In the proceedings of *UseR! 2014 - the International R User Conference*.

Baker, Monya. 2016. "Is There a Reproducibility Crisis?" *Nature* 533:452-454.

Berger, James. 2006. "The Case for Objective Bayesian Analysis." *Bayesian analysis* 1:385-402.

Bradburn, Michael J., Jonathan J. Deeks, Jesse A. Berlin, and A. Russell Localio. 2007. "Much Ado about Nothing: A Comparison of the Performance of Meta-analytical Methods with Rare Events." *Statistics in Medicine* 26:53-77.

Brown, Lawrence D., T. Tony Cai, and Anirban DasGupta. 2001. “Interval Estimation for a Binomial Proportion.” *Statistical Science* 16:101-117.

Button, Katherine S., John PA Ioannidis, Claire Mokrysz, Brian A. Nosek, Jonathan Flint, Emma SJ Robinson, and Marcus R. Munafò. 2013. "Power Failure: Why Small Sample Size Undermines the Reliability of Neuroscience." *Nature Reviews Neuroscience* 14:365-376.

Camerer, Colin F., Anna Dreber, Eskil Forsell, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, et al. 2016. “Evaluating Replicability of Laboratory Experiments in Economics. *Science* 351:1433-1436.

Cohen, Jacob. 1988. *Statistical Power Analysis for the Behavioral Sciences*. 2nd ed. Hove: Lawrence Erlbaum Associates.

Cohen, Jacob. 1992. “A Power Primer.” *Psychological Bulletin* 112:155-159.

Cohen, Jacob and Patricia Cohen. 1983. Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences. *Hillsdale, NJ: Erlbaum.*

Colquhoun, David. 2014. "An Investigation of the False Discovery Rate and the

Misinterpretation of P-values." *Royal Society Open Science* 1:1-15.

Cumming, Geoff. 2014. "The New Statistics: Why and How." *Psychological Science* 25:7-29.

Davenport, Ernest C. Jr. and Nader A. El-Sanhurry. 1991. “Phi/phimax: Review and Synthesis.” *Educational and Psychological Measurement* 51:821-828.

Dienes, Zoltan. 2014. "Using Bayes to Get the Most Out of Non-significant Results." *Frontiers in Psychology* 5:1-17.

Errington, Timothy M., Elizabeth Iorns, William Gunn, Fraser Elisabeth Tan, Joelle Lomax, and Brian A. Nosek. 2014. “An Open Investigation of the Reproducibility of Cancer Biology Research.” *Elife* 3:e04333.

Fanelli, Daniele. 2012. "Negative Results are Disappearing from Most Disciplines and Countries.” *Scientometrics* 90:891-904.

Ferguson, Christopher J. and Moritz Heene. 2012. “A Vast Graveyard of Undead Theories: Publication Bias and Psychological Science’s Aversion to the Null.” *Perspectives on Psychological Science* 7:555-561

Fiedler, Klaus, Florian Kutzner, and Joachim I. Krueger. 2012. "The Long Way from α-error Control to Validity Proper: Problems with a Short-sighted False-positive Debate." *Perspectives on Psychological Science* 7:661-669.

Fraley, Chris R. and Simine Vazire. 2014. "The N-pact Factor: Evaluating the Quality of Empirical Journals with Respect to Sample Size and Statistical Power." *PloS One* 9:1-12.

Franco, Annie, Neil Malhotra, and Gabor Simonovits. 2014. "Publication Bias in the Social Sciences: Unlocking the File Drawer." *Science* 345:1502-1505.

Freese, Jeremy. 2007a. “Overcoming Objections to Open-source Social Science.” *Sociological Methods & Research*, *36*(2), 220-226.

Freese, Jeremy. 2007b. “Replication Standards in Quantitative Social Science: Why Not Sociology?” *Sociological Methods & Research, 36*(2), 153-172.

Freese, Jeremy and David Peterson. 2017. “Replication in Social Science.” *Annual Review of Sociology* 43:147-165.

Gelman, Andrew and John Carlin. 2014. "Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors." *Perspectives on Psychological Science* 9:641-651.

Gelman, Andrew, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. 2013. *Bayesian Data Analysis.* 3rd ed. Boca Raton, FL: CRC Press.

Gelman, Andrew, Aleks Jakulin, Maria Grazia Pittau, and Yu-Sung Su. 2008. "A Weakly Informative Default Prior Distribution for Logistic and Other Regression Models." *The Annals of Applied Statistics* 2:1360-1383.

Gelman Andrew and Eric Loken. 2014. “The Statistical Crisis in Science: Data-dependent Analysis—a ‘Garden of Forking Paths’—Explains Why Many Statistically Significant Comparisons Don’t Hold Up.” *American Scientist* 102:460–465. Erratum at http://andrewgelman.com/2014/10/14/didnt-say-part-2/

Gelman, Andrew, Daniel Simpson, and Michael Betancourt. 2017. “The Prior Can Generally Only Be Understood in the Context of the Likelihood.” *arXiv preprint:1708.07487*. Available at: https://arxiv.org/abs/1708.07487

Ghosh, Malay. 2011. "Objective Priors: An Introduction for Frequentists." *Statistical Science* 26:187-202.

Gill, Jeff. 1999. “The Insignificance of Null Hypothesis Significance Testing.” *Political Research Quarterly* 52:647-674.

Glasman, Laura R. and Dolores Albarracín. 2006. “Forming Attitudes that Predict Future

Behavior: A Meta-analysis of the Attitude-Behavior Relation.” *Psychological Bulletin* 132:778-822.

Goodchild van Hilten, Lucy. 2015. "Why It’s Time to Publish Research ‘Failures’." *Science Communication*. Available at: https://www.elsevier.com/connect/scientists-we-want-your-negative-results-too.

Greenland, Sander. 2011. "Null Misinterpretation in Statistical Testing and Its Impact on Health Risk Assessment." *Preventive Medicine* 53:225-228.

Greenland, Sander. 2012. "Nonsignificance Plus High Power Does Not Imply Support for the Null Over the Alternative." *Annals of Epidemiology* 22:364-368.

Greenland, Sander, Stephen J. Senn, Kenneth J. Rothman, John B. Carlin, Charles Poole, Steven N. Goodman, and Douglas G. Altman. 2016. "Statistical Tests, P Values, Confidence Intervals, and Power: A Guide to Misinterpretations." *European Journal of Epidemiology* 31:337-350.

Grimes, David Robert, Chris T. Bauch, and John P.A. Ioannidis. 2018. "Modelling Science Trustworthiness Under Publish or Perish Pressure." *Royal Society Open Science* 5:1-14.

Hoenig, John M. and Dennis M. Heisey. 2001. "The Abuse of Power: The Pervasive Fallacy of Power Calculations for Data Analysis." *The American Statistician* 55:19-24.

Ioannidis, John P. 2005. “Why Most Published Research Findings are False.” *PLoS Med* 2:696-701.

Jasny, Barbara R., Gilbert Chin, Lisa Chong, and Sacha Vignieri. 2011. “Again, and Again, and Again….” *Science* 334:1225.

Jerolmack, Colin and Shamus Khan. 2014. “Talk is Cheap: Ethnography and the Attitudinal Fallacy.” *Sociological Methods and Research* 43:178-209.

Kaltenhauser, Jerome and Yuk Lee. 1976. “Correlation Coefficients for Binary Data in Factor Analysis.” *Geographical Analysis* 8:305-313.

Kim, Min-Sun and John E. Hunter. 1993. “Attitude-Behavior Relations: A Meta-analysis of Attitudinal Relevance and Topic.” *Journal of Communication* 43:101-142.

King, Gary. 1995. “Replication, Replication.” *PS: Political Science & Politics* 28:444-452.

King, Gary and Langche Zeng. 2001. "Logistic Regression in Rare Events Data." *Political Analysis* 9:137-163.

Kraus, Stephen J. 1995. “Attitudes and the Prediction of Behavior: A Meta-Analysis of the Empirical Literature.” *Personality and Social Psychology Bulletin* 21:58-75.

Kruschke, John K. 2011. "Bayesian Assessment of Null Values Via Parameter Estimation and Model Comparison." *Perspectives on Psychological Science* 6:299-312.

Kruschke, John K. 2015. "Doing Bayesian Data Analysis: A Tutorial Introduction with R, JAGS and Stan”

Kruschke, John K. and Torrin M. Liddell. 2018a. "The Bayesian New Statistics: Hypothesis Testing, Estimation, Meta-analysis, and Power Analysis from a Bayesian Perspective.” *Psychonomic Bulletin & Review* 25*:*178-206.

Kruschke, John K., and Torrin M. Liddell. 2018b. “Bayesian data analysis for newcomers.” *Psychonomic bulletin & review* 25:155-177.

Lakens, Daniël and Ellen R.K. Evers. 2014. "Sailing from the Seas of Chaos into the Corridor of Stability: Practical Recommendations to Increase the Informational Value of Studies." *Perspectives on Psychological Science* 9:278-292.

Ma, Yan, Haitao Chu, and Madhu Mazumdar. 2016. "Meta-analysis of Proportions of Rare Events–A Comparison of Exact Likelihood Methods with Robust Variance

Estimation." *Communications in Statistics-Simulation and Computation* 45:3036-3052.

Maxwell, Scott E., Michael Y. Lau, and George S. Howard. "Is Psychology Suffering from a Replication Crisis? What Does 'Failure to Replicate' Really Mean?" *American Psychologist* 70:487-498.

McDonald, Patrick. 2017. *Improving our Understanding of Employer Decision-making Thanks to Factorial Survey Analysis*. LIVES Working paper, 2017/61. Available at: http://dx.doi.org/10.12682/lives.2296-1658.2017.61

Murtaugh, Paul A. 2014. "In Defense of P Values." *Ecology* 95:611-617.

Muthén, Bengt. 2010. "Bayesian Analysis in Mplus: A Brief Introduction." *Unpublished manuscript. www. statmodel. com/download/IntroBayesVersion* 203.

O'Keefe, Daniel J. 2007. "Brief report: Post Hoc Power, Observed Power, a Priori Power, Retrospective Power, Prospective Power, Achieved Power: Sorting Out Appropriate Uses of Statistical Power Analyses." *Communication Methods and Measures* 1:291-299.

Oakes, Lisa M. 2017. "Sample Size, Statistical Power, and False Conclusions in Infant Looking Time Research." *Infancy* 22:436-469.

Occhiuto, Nicholas. 2017. "Investing in Independent Contract Work: The Significance of Schedule Control for Taxi Drivers." *Work and Occupations* 44:268-295.

Olivier, Jake and Melanie L. Bell. 2013. "Effect Sizes for 2×2 Contingency Tables." *PLoS One* 8:1-7.

Open Science Collaboration. 2015. "Estimating the Reproducibility of Psychological Science. *Science* 349:943.

Pager, Devah. 2003. "The Mark of a Criminal Record." *American Journal of Sociology* 108:937-975.

Pager, Devah. 2002. *The Mark of a Criminal Record* (PhD), University of Wisconsin-Madison, Madison, WI.

Pager, Devah and Lincoln Quillian. 2005. “Walking the Talk? What Employers Say versus What They Do.” *American Sociological Review* 70:355-380.

Project TIER Teaching Integrity in Empirical Research. Retrieved April 15, 2018 (https://www.projecttier.org/).

Reich, Suzanne E. 2017. “An Exception to the Rule: Belief in Redeemability, Desistance Signals, and the Employer’s Decision to Hire a Job Applicant with a Criminal Record.” *Journal of Offender Rehabilitation* 56:110-136.

Schoot, Rens, David Kaplan, Jaap Denissen, Jens B. Asendorpf, Franz J. Neyer, and Marcel AG Aken. 2014. "A Gentle Introduction to Bayesian Analysis: Applications to Developmental Research." *Child Development* 85:842-860.

Schuman, Howard and Michael P. Johnson. 1976. “Attitudes and Behavior.” *Annual Review of Sociology* 2:161-207.

Simmons, Joseph P., Leif D. Nelson, and Uri Simonsohn. 2011. “False-positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science, 22*:1359-1366.

Stokes, Maura, Fang Chen, and Funda Gunes. 2014. "An Introduction to Bayesian Analysis with SAS/STAT® Software." In *Proceedings of the SAS Global Forum 2014 Conference*, SAS Institute Inc, Cary, USA. Available at: https://support.sas.com/resources/papers/proceedings14/SAS400-2014.pdf

Tittle, Charles R. 2004. "The Arrogance of Public Sociology." *Social Forces* 82:1639-1643.

Turner, Rebecca M., Sheila M. Bird, and Julian PT Higgins. 2013. "The Impact of Study Size on

Meta-analyses: Examination of Underpowered Studies in Cochrane Reviews." *PLoS One* 8:1-8..

Vaisey, Stephen. 2014. “The ‘Attitudinal Fallacy’ is a Fallacy: Why We Need Many Methods to Study Culture.” *Sociological Methods & Research* 43:227-231.

Vadillo, Miguel A., Emmanouil Konstantinidis, and David R. Sharks. 2016. “Underpowered Samples, False Negatives, and Unconcious Learning.” *Psychonomic Bulletin & Review* 23:87-102.

Vuolo, Mike, Christopher Uggen, and Sarah Lageson. 2016. "Statistical Power in Experimental Audit Studies: Cautions and Calculations for Matched Tests with Nominal Outcomes.” *Sociological Methods & Research* 45:260-303.

Vasishth, Shravan and Andrew Gelman. 2017. "The Illusion of Power: How the Statistical Significance Filter Leads to Overconfident Expectations of Replicability." *arXiv preprint arXiv:1702.00556*.

Wasserstein, Ronald L. and Nicole A. Lazar. 2016. "The ASA's Statement on P-values: Context, Process, and Purpose." *The American Statistician* 70:129-133.

Western, Bruce. 1999. "Bayesian analysis for sociologists: An introduction." *Sociological Methods & Research* 28:7-34.

TABLE 1. Counterfactual Associations between What Employers Say in a Vignette and What They Do in an Audit, by Expected Effect Magnitude and Vignette Coding Condition

	1. "Maximum"		2. "Strong"		3. "Moderate"	
	Expected Audit Behavior:		Expected Audit Behavior:		Expected Audit Behavior:	
Observed Survey Group:	**No Callback**	**Callback**	**No Callback**	**Callback**	**No Callback**	**Callback**
Panel 1A. (N=156)						
Somewhat/very likely to hire	83 (.86)	13 (.14)	85 (.89)	11 (.11)	86 (.90)	10 (.10)
Somewhat/very unlikely to hire	60 (1.0)	0 (.00)	58 (.97)	2 (.03)	57 (.95)	3 (.05)
Φ (Φ/Φ_{max})	.24 (1)		.14 (.60)		.10 (.40)	
Proportion Diff. [Odds Ratio]	*.14 [∞]*		*.08 [3.75]*		*.05 [2.21]*	
Power of Fisher's Exact Test	.99		.37		.14	
Panel 1B. (N=156)						
Very likely to hire	17 (.77)	5 (.23)	19 (.86)	3 (.14)	20 (.91)	2 (.09)
At most somewhat likely	134 (1.0)	0 (.00)	132 (.99)	2 (.01)	131 (.98)	3 (.02)
Φ (Φ/Φ_{max})	.45 (1)		.24 (.53)		.14 (.30)	
Proportion Diff. [Odds Ratio]	*.23 [∞]*		*.12 [10.42]*		*.07 [4.37]*	
Power of Fisher's Exact Test	.98		.66		.28	
Panel 1C. (N=156)						
At least somewhat unlikely	108 (.87)	16 (.13)	109 (.88)	15 (.12)	110 (.89)	14 (.11)
Very unlikely to hire	32 (1.0)	0 (.00)	31 (.97)	1 (.03)	30 (.94)	2 (.06)
Φ (Φ/Φ_{max})	.17 (1)		.12 (.70)		.07 (.39)	
Proportion Diff. [Odds Ratio]	*.13 [∞]*		*.09 [4.27]*		*.05 [1.91]*	
Power of Fisher's Exact Test	.66		.22		.06	
Panel 1D. (N=54)						
Very likely to hire	17 (.77)	5 (.23)	18 (.82)	4 (.18)	19 (.86)	3 (.14)
Very unlikely to hire	32 (1.0)	0 (.00)	31 (.97)	1 (.03)	30 (.94)	2 (.06)
Φ (Φ/Φ_{max})	.39 (1)		.26 (.66)		.13 (.33)	
Proportion Diff. [Odds Ratio]	*.23 [∞]*		*.15 [6.89]*		*.07 [2.37]*	
Power of Fisher's Exact Test	.78		.37		.11	

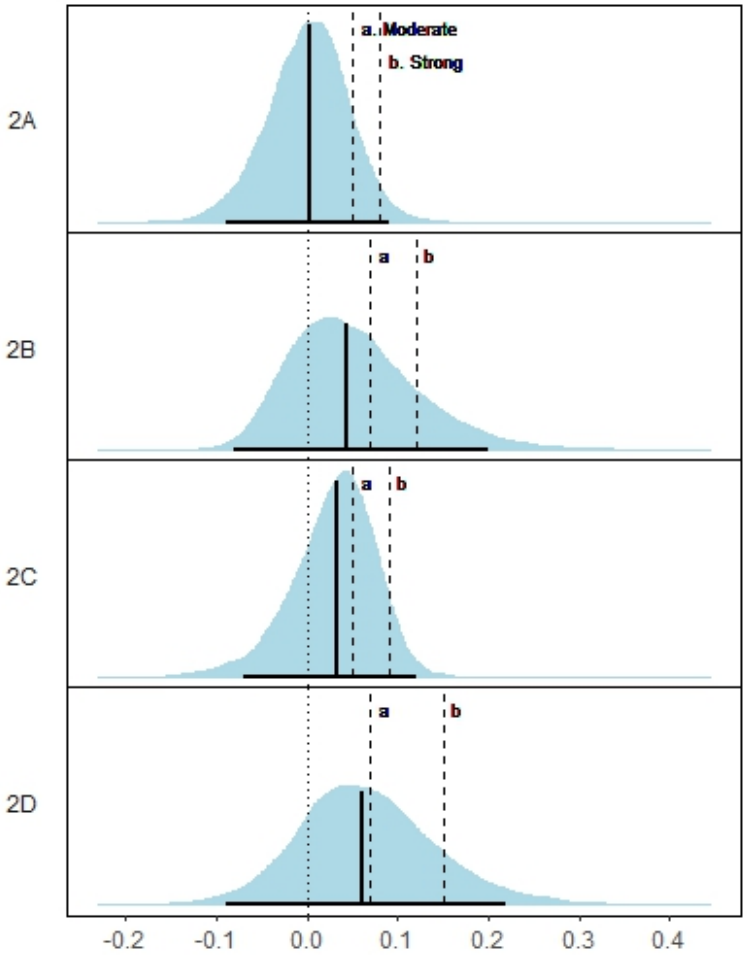
Notes: Joint frequency distributions are counterfactual "expected" audit callbacks of applicants with criminal records by employers' reported willingness to hire such candidates in a survey vignette. Distributions were calculated by starting with the marginal frequencies for employers' willingness to hire candidates with criminal records in the overlapping vignette/audit sample (Pager and Quillian 2005:377). Next, expected callback rates for employers in the "more likely" group (top row in each panel) were calculated by multiplying *very likely* and *somewhat likely/unlikely* subgroup frequencies, respectively, by observed callback probabilities for applicants *without* and *with* criminal records (.226 and .103; see Pager 2003). "Maximum possible" associations between vignette reports and audit behaviors (Column A) were then obtained by fixing the "less likely" callback rate to zero. Strong and moderate associations (Columns B and C) were calculated using mean meta-analytic attitude-behavior correlations as benchmarks (i.e., Φ/Φ_{max} values comparable to $r \geq .52$ and $r \approx .38$).

TABLE 2. Observed Associations between What Employers Say in a Vignette and What They Do in an Audit When Considering Job Applicants with Criminal Records

Survey responses:	(1) Audit: No Callback	(2) Audit: Callback	(3) *Odds Ratio* Exact Test	(4) Rel. Freq. [b] [95% HDI]	(5) *Est. Diff.* [c] [95% HDI]	(6) $P_{1>2}$ [d]
Panel 2A. (N=156)	***Observed*** [a]					
Somewhat/very likely to hire	89 (.927)	7 (.073)	*OR=1.10* *p*≤1.00	.08 [.03, .14]	*0* [-.09, .08]	***$P_{1>2}$=.52***
Somewhat/very unlikely to hire	56 (.933)	4 (.067)		.08 [.02, .15]		
Panel 2B. (N=156)	***Observed*** [a]					
Very likely to hire	20 (.909)	2 (.091)	*OR=1.39* *p*=.655	.11 [.02, .25]	*.04* [-.07, .19]	***$P_{1>2}$=.76***
At most somewhat likely	125 (.933)	9 (.067)		.07 [.03, .12]		
Panel 2C. (N=156)	***Observed*** [a]					
At least somewhat unlikely	114 (.870)	10 (.130)	*OR=2.71* *p*=.463	.09 [.04, .14]	*.03* [-.07, .11]	***$P_{1>2}$=.76***
Very unlikely to hire	31 (.969)	1 (.031)		.05 [.00, .14]		
Panel 2D. (N=54)	***Observed*** [a]					
Very likely to hire	20 (.909)	2 (.091)	*OR=3.03* *p*=.560	.12 [.02, .25]	*.06* [-.08, .23]	***$P_{1>2}$=.81***
Very unlikely to hire	31 (.969)	1 (.031)		.05 [.00, .14]		

[a] "*Observed*" frequency distributions (2A–2D) calculated from Pager and Quillian's Appendix (2005:377).
[b] "Rel. Freq." refers to the estimated relative frequency of success, or the median of the posterior distribution for employers in the "more likely" (θ_1) or "less likely" (θ_2) groups. 95% HDI refers to the 95% highest density interval.
[c] "*Est. Diff.*" is the estimated difference in callback proportions between employers in the "more likely" and "less likely" groups, or the difference in medians of the posterior distributions for "more likely" and "less likely" groups (θ_1–θ_2).
[e] "$P_{1>2}$" is the estimated probability that the relative callback frequency is larger for employers in "more likely" group compared to the "less likely" group, or the proportion of the posterior density for (θ_1–θ_2) that is greater than zero.

FIGURE 1. Distributions of Credible Effect Sizes for the Employer Vignette/Audit Overlap by Vignette Coding Condition, Contrasted with Counterfactual Moderate and Strong Expected Effect Sizes



Design/measurement variations and results

Proportion difference in callbacks: *somewhat/very likely* vs. *somewhat/very unlikely*

Probability that the difference between "more" and "less" willing group is:
… greater than zero ($\theta_1 - \theta_2 > 0$; $OR > 1.0$) = .52
… equal to/greater than the moderate counterfactual ($\theta_1 - \theta_2 \geq .05$; $OR \geq 2.21$) = .13
… equal to/greater than the strong counterfactual ($\theta_1 - \theta_2 \geq .08$; $OR \geq 3.75$) = .03

Proportion difference in callbacks: *very likely* vs. *at most somewhat likely*

Probability that the difference between "more" and "less" willing group is:
… greater than zero ($\theta_1 - \theta_2 > 0$; $OR > 1.0$) = .75
… equal to/greater than the moderate counterfactual ($\theta_1 - \theta_2 \geq .07$; $OR \geq 4.37$) = .34
… equal to/greater than the strong counterfactual ($\theta_1 - \theta_2 \geq .12$; $OR \geq 10.42$) = .16

Proportion difference in callbacks: *at least somewhat unlikely* vs. *very unlikely*

Probability that the difference between "more" and "less" willing group is:
… greater than zero ($\theta_1 - \theta_2 > 0$; $OR > 1.0$) = .76
… equal to/greater than the moderate counterfactual ($\theta_1 - \theta_2 \geq .05$; $OR \geq 1.91$) = .34
… equal to/greater than the strong counterfactual ($\theta_1 - \theta_2 \geq .09$; $OR > 4.27$) = .07

Proportion difference in callbacks: *very likely* vs. *very unlikely*

Probability that the difference between "more" and "less" willing groups is:
… greater than zero ($\theta_1 - \theta_2 > 0$; $OR > 1.0$) = .81
… equal to/greater than the moderate counterfactual (a; $\theta_1 - \theta_2 \geq .07$; $OR \geq 2.37$) = .45
… equal to/greater than the strong counterfactual (b; $\theta_1 - \theta_2 \geq .15$; $OR > 6.89$) = .14

Note: Blue shaded area displays probability density curve for estimated difference in audit callback proportions between employers who are "more willing" versus "less willing" to hire a hypothetical candidate with a criminal record ($\theta_1 - \theta_2$; see Column 5 in Table 2). Solid vertical line is the median of posterior density. Solid horizontal line is the 95% credible interval (i.e., highest density interval or HDI). Dotted vertical line (at "0") represents "no association" between employer vignette and audit, while dashed vertical lines reflect (a) "moderate" and (b) "strong" counterfactual expected associations (see Table 1).

ENDNOTES

[i] We refer readers interested in calculating such errors to their original paper. For potent discussions of the perils of power, see Hoenig and Heisey (2001) and Greenland (2012). Subsequently, we recommend Bayesian analysis with default priors, which can also be used to estimate sign and magnitude errors (see Gelman and Carlin 2014, ftn. 4).

[ii] For other accessible introductions to Bayesian concepts and logic, see Western (1999) and Kruschke and Liddell (2018b).

[iii] For more information, see https://github.com/rasmusab/bayesian_first_aid/blob/master/README.md.

[iv] See: https://www.ibm.com/support/knowledgecenter/SSLVMB_25.0.0/statistics_mainhelp_ddita/spss/advanced/idh_bayesian.html

[v] See: https://www.stata.com/new-in-stata/bayes-prefix/

[vi] The first author of this manuscript contacted Devah Pager via email in August 2017 to request direct access of the initial Milwaukee audit data or indirect access via sharing through a public repository. Pager responded immediately and offered to attempt to track down the original data. The original audit data were not made available as of the date of submission of this manuscript.

[vii] Refer to: https://cran.r-project.org/web/packages/Exact/Exact.pdf

[viii] For a related discussion, refer to Cohen and Cohen's (1983:65-67) discussion of how the distributions of X and Y affect the size of correlation coefficients. Also, note that similar issues arise in interpreting alternative effect size measures, such as odds ratios, in the presence of rare events (see Olivier & Bell 2013).

[ix] Recall, these "maximum possible" distributions do not represent *plausible* effects; rather, they are more appropriately considered an upper-bound for plausible effect sizes.

[x] Refer to: http://www.sumsar.net/blog/2014/01/bayesian-first-aid-binomial-test/.

[xi] Our general conclusions are also robust to alternative specification of Jeffreys non-informative prior distribution (see Brown, Cai, and DasGupta, 2001; Ghosh, 2011).