ISSN 1936-5330

kc FED

The Effect of the Conservation Reserve Program on Rural Economies: Deriving a Statistical Verdict from a Null Finding

Jason P. Brown, Dayton M. Lambert and
Timothy R. Wojan

May 2018
RWP 18-04
https://dx.doi.org/10.18651/RWP2018-04

FEDERAL RESERVE BANK *of* KANSAS CITY

The Effect of the Conservation Reserve Program on Rural Economies: Deriving a Statistical Verdict from a Null Finding*

Jason P. Brown†, Dayton M. Lambert‡, Timothy R. Wojan§

May 21, 2018

Abstract

This article suggests two methods for deriving a statistical verdict from a null finding, allowing economists to more confidently conclude when "not significant" can in fact be interpreted as "no substantive effect." The proposed methodology can be extended to a variety of empirical contexts where size and power matter. The example used to demonstrate the method is the Economic Research Service's 2004 Report to Congress that was charged with statistically identifying any unintended negative employment consequences of the Conservation Reserve Program (the Program). The report failed to identify a statistically significant negative long-term effect of the Program on employment growth, but the authors correctly cautioned that the verdict of "no negative employment effect" was only valid if the econometric test was statistically powerful. We replicate the 2004 analysis and use new methods of statistical inference to resolve the two critical deficiencies that preclude estimation of statistical power by economists: 1) positing a compelling effect size, and 2) providing an estimate of the variability of an unobserved alternative distribution using simulation methods. We conclude that the test used in the report had high power for detecting employment effects of -1 percent or lower resulting from the Program, equivalent to job losses reducing a conservative estimate of environmental benefits by a third.

Keywords: power analysis, Monte Carlo simulation, hypothesis testing

JEL Classification Numbers: C12, Q42, R11

*We thank Terrance Hurley, John Pender, Daniel Hellerstein, David McGranahan, Patrick Sullivan, and two anonymous reviewers for their constructive comments and Colton Tousey for excellent research assistance. Lambert's research was supported by USDA Hatch Multistate Project NE-1749 funding. The views expressed are those of the authors and are not attributable to the Federal Reserve Bank of Kansas City, the Federal Reserve System, the Economic Research Service or the U.S. Department of Agriculture. Any remaining errors are those of the authors.

†Federal Reserve Bank of Kansas City, jason.brown@kc.frb.org

‡University of Tennessee Institute of Agriculture, Department of Agricultural & Resource Economics, dlamber1@utk.edu

§USDA, Economic Research Service, twojan@ers.usda.gov

1 Introduction

Researchers often report results that are not statistically significant, also known as null findings. But an insignificant result does not necessarily mean an unimportant one. The objective of this article is to suggest a method for deriving probabilities for null findings, allowing economists to more confidently conclude when "not significant" can in fact be interpreted as "no substantive effect." The example used to demonstrate the method is the Economic Research Service's (ERS) 2004 Report to Congress on the economic implications of the Conservation Reserve Program (CRP).

Continued employment and population decline in farm-dependent counties through the 1990s raised concern that agricultural programs encouraging the removal of environmentally vulnerable land from production might have cost jobs. The ERS study did identify worse employment growth performance in farm-dependent counties with high-CRP enrollments compared to low-CRP peers. However, the analysis was unable to attribute lost employment to CRP enrollments. The combination of multiple model specifications that failed to find statistically significant negative employment impacts of CRP supported a cautious conclusion of "*no evidence of negative employment impacts from CRP.*" However, the report correctly noted that the "*absence of evidence is not evidence of absence.*" The statistical power of the test was unknown. The authors correctly cautioned that there was no unequivocal statistical evidence that "not significant" *could be* interpreted as "no negative effect."

Estimating the statistical power that was unknown in the 2004 report requires addressing two critical deficiencies characterizing the majority of econometric studies using null hypothesis statistical testing (NHST): 1) positing *a priori* a compelling effect size (i.e., the minimum effect considered economically significant), and 2) providing an approximation of the shape, location, and scale of an unobserved alternative distribution. The first deficiency is filled through back-of-the-envelope calculations equating program costs to program benefits. These ballpark estimates provide a conceivable range of adverse employment effects following enrollment of cropland into the CRP. The second deficiency is conceptually and

computationally more challenging. We develop candidate alternative distributions using two simulation approaches: 1) a bootstrap resampling procedure and 2) a Bayesian approach forcing an effect size with a strong prior.

Power estimates from the 2004 Report are challenging from the standpoint of both conventional practice and the explicit charge from Congress to search for "any effect". Our findings suggest that the tests used to search for "any effect" were low-power. A strict reading of the Congressional charge and of the NHST protocol would require suspending judgment on the likely effect of CRP on employment growth. If the *de facto* charge was to search for an economically significant effect of CRP on employment – i.e., an effect that an informed person would not regard as miniscule – then our replication reinforces the original findings. Since the test to detect a negative effect of -1 percent was powerful, the null finding can be interpreted as "no economically significant effect." Lower than this posited effect size would require suspension of judgment. The broader implications for econometric practice are discussed in the conclusion.

2 The Challenge of Relying on NHST to Inform Policy

The two dominant ways of using statistical analysis are either as an instrument of scientific exploration or as an instrument to aid decision-making. The work of Ronald Fisher provides the foundation for the former. The protocol developed by Jerzy Neyman and Egon Pearson provides the foundation for the latter (Christensen, 2005).

The key construct underlying Fisherian NHST is that scientific exploration begins from a position of ignorance. Compelling alternative hypotheses are unknown. The benefits this approach provides are immediate: 1) only a single distribution is required for testing if an estimate is statistically different from the presumed null; 2) the parameters of the null distribution are derived solely from sample data with no requirement for prior or auxiliary information; and 3) in the case of a statistically significant result, the protocol provides a

measure of confidence in that verdict. The major cost of this method is that no statistical inference is possible for nonsignificant results. The proof by contradiction has failed. The only valid verdict is to suspend judgment. In its purest sense, this cost in scientific exploration is zero because nonsignificant findings carry no normative implications.

The dominant frequentist alternative to Fisherian NHST is the Neyman-Pearson protocol that was developed explicitly as a statistical tool to aid decision-making (Tweeten, 1983). Within this framework, the researcher is required to collect information not available in the sample. The researcher must arrive at a relevant effect size that defines the mean of the alternative distribution. Relevance might be derived in a number of ways including predictions from theory, results from computational models, or a breakeven point for a treatment or policy. The research must also posit what the alternative distribution looks like in terms of its location, dispersion, or shape. These features could be informed by a literature review. With this information, the researcher can conduct an *ex ante* power analysis to determine the sample size needed to produce a powerful test. The upfront costs of this approach produce their benefits at the end of the analysis when the findings are used to inform a decision. The verdict from a statistically significant result parallels that in NHST, but the verdict from a nonsignificant result is also informative: "*with X level of confidence, the absolute magnitude of the treatment effect fails to meet the minimum* <posited effect size> *required for economic significance.*"

The Neyman-Pearson protocol is fairly common in applied statistical disciplines where equivocal findings could result in significant monetary costs, such as biomedical research. The most persuasive explanation for why econometrics has placed more emphasis on Fisherian NHST versus Neyman-Pearson is the much greater difficulty of positing an alternative distribution (Wojan et al., 2014). Lacking conjectures for an unobserved alternative distribution, it made little sense to incorporate effect size into an econometric analysis because an explicit estimate of statistical power would be impossible. Instead, econometrics grafted the concepts of statistical power and Type II error (falsely failing to reject the null hypoth-

esis) onto the Fisherian NHST protocol as tertiary issues of concern that would rarely be considered.

Our conjecture is that simple (often implicit) rules of thumb and heuristics are what allow economists to apply NHST as a statistical aid to decision-making without discarding every nonsignificant finding as uninformative, as a strict Fisherian would require. The simplest determinant of statistical power is sample size and so the problem of ensuring tests of adequate power is often reduced to ensuring tests of adequate sample size. There are no hard and fast rules, but an appreciation of "adequate sample size" is something economists develop through experience. If an analysis of a given sample size produces statistically significant results, then tests for other specifications and dependent variables in the analysis may be assumed to be adequately powerful to conclude that a program had no effect for some outcomes, without an actual estimation of the statistical power. Unfortunately, these approaches that abstract from effect size and treatment variability – the more complex determinants of statistical power – also abstract from the most powerful determinants of statistical power. For example, if the effect size that matters is "large", a relatively small sample may provide a very powerful test. Conversely, if the variation around the treatment effect is large, then a "reasonably large dataset" may only provide weak tests.

While academic economists express confidence that the scholarly community can effectively regulate their NHST-hybrid to guard against erroneous statistical inference (Hoover and Siegler, 2008), the American Statistical Association recently expressed renewed concern over adequacy of statistical significance and p-values for informing decisions (Wasserstein, 2016). The three principles most germane to economists doing policy relevant research are:

> ...
> *3. Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.*
> *4. Proper inference requires full reporting and transparency.*
> *5. A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.*

Professional opinion regarding the adequacy of statistical power would appear to fall short

of the requirement for "full reporting and transparency."

Because the statistical power of an empirical test is objective information, the most informative studies provide power estimates whenever a nonsignificant finding is relevant to a public policy question (see Nickerson et al. 2017, who suggest providing statistical power assessments as a best practice for Federal program evaluation activities). However, *ex ante* statistical power assessments may often be infeasible due to the novelty inherent in evaluating new programs or initiatives (Wojan et al., 2014). For transparency, the 2004 CRP Report included the caveat that the statistical power of the econometric test was unknown (Sullivan et al., 2004). The flexibility for conducting analyses and making inferences even when an exemplary dataset and prior studies are not available is made explicit in Practice 4 of the *Principles and Practices of a Federal Statistical Agency* (National Academies of Science, Engineering, and Medicine, 2017, p. 81) that addresses openness "about the strengths and limitations of its data." The question is whether the assumed infeasibility of statistical power assessments is truly binding. To answer this question, we develop approaches to generate *ex post* statistical power estimates to supplement the interpretation of nonsignificant findings. The first approach uses a bootstrap resampling-with-replacement procedure. The second approach is Bayesian, estimating power based on posterior marginal distributions of posited effect sizes.

3 Congressionally Mandated Study of the Conservation Reserve Program's Economic Implications

The Conservation Reserve Program was authorized in 1985 for the purpose of providing public benefits by taking environmentally vulnerable agricultural land out of production. The CRP had an acreage cap and farmers submitted bids, ensuring that the benefits provided were secured at a reasonable cost to the government. If proceeds from these contracts went strictly to farmers, the program may be expected to have limited negative effect on

the economic activity in farm dependent counties. However, since the proceeds went to landowners, who may or may not have resided in the county, there was the possibility the local economic losses from the decline in agricultural production would not be fully compensated by CRP payments. And since many counties with relatively more CRP contracts appeared to be losing jobs and population during a period of national prosperity, the concern was that taking agricultural land out of production might be exacerbating the problem. Thus, Congress requested a study from ERS to examine the economic implications of the program.

The 2004 ERS study provides a comprehensive examination of the effect of CRP on farm and non-farm rural economies including discussions of CRP rental payments and absentee landowners, the environmental and scenic impacts of CRP, and the anticipated upstream effects of CRP on businesses providing inputs to farming (Sullivan et al., 2004). The comprehensiveness of the report reinforces the story that the statistical analysis of employment trends supported but could not definitively confirm: i.e., that implementation of the CRP had small negative short term impacts on farm-dependent counties with high CRP enrollments but these impacts were not evident in the longer term due perhaps to observed increases in recreational spending. The NHST conundrum of not knowing whether the nonsignificant estimate of high-CRP enrollment on long-term employment growth could be interpreted as "no effect" or should be interpreted as a weak test is what the present study helps to resolve.

Testing for the effects of high-CRP enrollment on employment growth presented the challenge of adequately controlling for endogenous selection. The assumption from the outset was that many of the conditions that would support high CRP enrollments were also conditions that would be associated with long-term employment decline. The research design that was eventually implemented used a quasi-experimental matched pair protocol using the Mahalanobis-metric procedure, matching individual high-CRP counties with similarly situated low-CRP counties. Conceptually, if paired counties were nearly identical in those attributes explaining employment growth and program participation, then any observed difference in employment growth would be attributable to differences in CRP enrollments.

Empirically, it turned out that significant differences between the treatment (high CRP) and control (low CRP) counties persisted even after optimal matching. A difference-in-difference (DID) specification was required to control for differences in matching variables that persisted to isolate the effect of high versus low CRP enrollments on job growth.[1].

One-hundred and ninety high-CRP counties were matched 1-to-1 with low-CRP counties. Table 1 provides information on the mean value of industrial, labor market, and farm structure variables for the two groups. The high-CRP counties tended to be more dependent on agriculture and government payments, had lower shares of employment in manufacturing, and were more likely to be located in the Great Plains. Had the matching algorithm found closer matches on these variables, then simply comparing the average employment growth across groups would have been informative of the impact of high-CRP enrollments. However, given differences in structural characteristics, it is reasonable to assume that many factors other than CRP enrollment contributed to differences in job gains. Table 2 demonstrates an employment change difference of 5.8 percent between high-CRP and matched counties.

The descriptive statistics from the matching exercise suggest the possibility that high-CRP enrollment may be strongly associated with poor employment growth performance. The critical question is whether any of this poorer performance is attributable to high-CRP enrollments.

The ERS researchers specified the DID regression equation to isolate the effect of high-CRP enrollment on employment growth, controlling for potentially confounding differences in other county attributes. Multiple specifications were estimated to guard against erroneous results due to misspecification error. Short-term regressions did find negative impacts of high-CRP on employment growth that were statistically significant in 7 of 20 alternative specifications (magnitudes of these estimates were not provided). The alternative specifications only suggested that the relatively small sample size of 190 matched pairs was adequately

[1]Two reviewers noted that recent developments in matching techniques would reduce the risk of a misspecified post-matching regression model using balanced matching procedures (Diamond and Sekhon, 2013) or a matching regression with adjustment (Abadie and Imbens, 2011). These approaches could also reduce variability that could be introduced during the matching procedure (Ho et al., 2007).

powerful. However, the main purpose of the alternative specifications was to increase confidence of researchers that the failure to produce statistically significant negative results in the long-run regressions were in fact informative. The 20 specifications estimated for the long-term dependent variable did produce one negative coefficient estimate that was not statistically significant, and 3 positive coefficient estimates significant at the 10% level.

The discussion in the report summarizing the implications of the regression exercise are a textbook demonstration of "provid[ing] objective information" (Principle 1) that recognized "limitations of the data" (Practice 4) outlined in the *Principles and Practices of a Federal Statistical Agency:*

> *Between the matched-pair and study data sets, the different measures of CRP usage, and other variations as discussed in Appendix A, we have 20 different estimates of the relationship between CRP use and population and employment trends. This approach allows us to assess the consistency of the matched-pair estimations. Given that estimated coefficients can change from one model to the next, consistent estimates provide some confidence that the absence of statistical significance can be interpreted as "CRP has no effect," even though we do not know the probability of a Type II or false negative error. Since the absence of evidence is not evidence of absence, this approach helps to corroborate the findings from the matched-pair analysis* (page 31).

In this discussion, the heuristic of robustness is used to reinforce the inference from a null finding. While robustness tests often provide valid checks of empirical findings, a challenge is that if a statistical test is in fact weak, numerous re-specifications will only provide additional evidence of weak power.

4 Deriving a Statistical Answer

As noted in the ERS report, deriving a statistical answer requires "know[ing] the probability of a Type II error or false negative error." Clearly, if the test had a high probability of detecting a negative effect of high-CRP enrollment on employment growth, then a non-significant finding could be interpreted as "CRP has no effect." Knowing the probability of a Type II error requires estimating the statistical power of the test, which requires in turn

positing an effect size that matters and producing a credible, though unobserved, alternative distribution.

The economics discipline has been slow to address the issue of positing relevant effects sizes. McCloskey and Ziliak (1996, p. 105) examined the issue and found that fewer than 30 percent of articles published in the *American Economic Review* in the 1980s discussed "the scientific conversation within which a coefficient would be considered large or small." So consideration of the magnitude of estimates was relatively rare even after estimates were available. Consideration of effect sizes prior to estimation was not examined explicitly by McCloskey and Ziliak, but the 4.4 percent of articles that had "consider[ed] the power of the test" may have done this. By the 1990s, eight percent of articles published in *AER* considered the power of the test suggesting a very modest improvement (Ziliak and McCloskey, 2004).

Congress charged ERS with identifying *any* negative impacts (Sullivan et al., 2004). Positing an effect size for the purpose of analysis could be interpreted as inconsistent with Congressional intent because the effect size that mattered was explicit: *any effect*. Theoretically, the power of the test will approach the size of the test as the effect size goes to zero, meaning that tests of any effect' will have low power. However, an objective, impartial resolution to the problem could be to provide a range of possible effect sizes, given a credible and transparent method of determining that range. If the magnitude of those effect sizes can be illuminated with a discussion of their economic relevance, then policymakers will have a much richer set of information guiding their normative decisions. Providing a range of effect sizes that might matter does not bias the analysis. The final decision regarding what matters is retained by the policymaker.

Describing a worst case scenario for unintended adverse effects of the CRP provides a compelling case of what would constitute an upper-bound effect size benchmarked to economic and environmental data. Equating job losses to the environmental benefits qualifies as such a scenario. Arriving at an approximation of this figure is all that is required. The number is not intended to inform policy but merely to provide a reference point.

Simplifying assumptions that allow a back-of-the-envelope derivation include: 1) program benefits are equivalent to direct program costs; 2) these program benefits can be allocated to the study as the share of program acres in treatment (high-CRP) counties; 3) there are "pure controls"; i.e., no CRP acres in low-CRP counties; and 4) one job in the year 2000 in treatment counties can be valued at $23,897. This value is the average earnings per nonmetropolitan job derived from the Bureau of Economic Analysis. Arriving at a ballpark employment loss percentage is calculated as the job equivalent cost (benefit) of the program (= "a") times the treatment county share of the program (b = program acres in treatment counties divided by total number of program acres), divided by total employment in the treatment counties (= "c"):

$$\text{Job Equivalent Cost of Program} \times \text{Treatment Counties Share of Program} \times [1/\text{Treatment County Jobs}] =$$
$$\frac{\$23.7 \text{ billion}}{\$23,897}(a) \times \frac{508,000 \text{ acres}}{33,981,000 \text{ acres}}(b) \times \frac{1}{537,398 \text{ jobs}}(c) = 2.76\% \quad (1)$$

These grossly simplified (though we believe reasonable) assumptions provide useful information for characterizing a reference point at which the program exerts adverse effects on employment in CRP counties. Remembering that employment growth in treatment counties lagged control counties by 5.8 percent, attributing half this loss to high-CRP enrollments would amount to a full negation of expected environmental benefits. If this worst case scenario was in fact supported by the analysis, then Congress could have a basis on economic efficiency grounds for modifying the program. However, adverse effects below this reference point could also provide an economic basis for modifying the CRP. Effect sizes roughly one half, third, or a fifth of the worst case scenario would correspond to an effect size of negative 1.5 percent, 1 percent and 0.5 percent, respectively. To assess the power of detecting "any effect", an effect size of negative 0.1 percent is also included in the *ex post* power estimates.

Arriving at a credible estimate of an unobserved alternative distribution is technically

more challenging than positing an effect size. And, unlike the effect size exercise, producing a range of distributions would misinterpret the function of the alternative distribution in a statistical power analysis; i.e., to provide an accurate representation of the phenomena of interest in the population. Traditionally, this has been done through extensive literature searches. However, the CRP study was the first of its kind and the conventional approach was impossible.

5 Ex Post Power Simulation with Monte Carlo Resampling

We replicate first the long-run local employment growth model used to evaluate CRP by Sullivan et al. (2004). The authors reported the CRP estimates for several models, but only complete results were reported for one version. We selected that specification to benchmark our *ex post* power analysis. In the Sullivan et al. study, employment growth between 1985 and 2000 was estimated using ordinary least squares (OLS) on the differenced values between matched pairs of high-CRP (HCRP) and low-CRP (LCRP) counties with the linear model:

$$y_i = \left[ln\left(\frac{emp_{i,2000}^{HCRP}}{emp_{i,1985}^{HCRP}}\right) - ln\left(\frac{emp_{i,2000}^{LCRP}}{emp_{i,1985}^{LCRP}}\right)\right] = X_i\beta + \varepsilon_i, \tag{2}$$

where i indexes a matched pair; $X_i = (X_i^{HCRP} - X_i^{LCRP})$ is an n by k matrix of matched pair differences in information on CRP payments and conditioning measures, including local socioeconomic and agricultural characteristics; and β is a k by 1 vector of coefficients (tables 1 and 2). The variable ε is independent and identically distributed random component with mean zero and a constant variance.

Equation 2 was estimated with the 190 matched pairs from the original study. The replicated OLS estimates are in Table 3. The results are nearly identical to those reported in Table A.3 of Sullivan et al. (2004). The variable of concern in this regression is the ratio

of county level payments from CRP over total income (X_{CRP}). We find a one standard deviation increase in the CRP to total income ratio would be associated with a positive and statistically significant 0.24 percent increase in employment growth.

The first step of the Monte Carlo resampling procedure entails selecting the value of $\bar{\beta}_{CRP}$; i.e., values indicating effect sizes under the alternative hypothesis. In this example, a range of effect sizes were evaluated, given the absence of a specific effect of interest in the charge from Congress. The alternative hypotheses of employment growth response to CRP were set to $\bar{\beta}_{CRP} = -0.001, -0.005, -0.010, -0.015, \text{and} -0.027$. These values correspond with the smallest effect size of -0.1 percent in employment growth from a unit change in CRP to the largest effect size of -2.7 percent in employment growth. Determining effect sizes can be subjective. However, in the current analysis we chose to work from our estimate of the job loss (-2.7 percent) that would offset CRP benefits and smaller effect sizes approaching zero (Equation 1). We specifically chose negative numbers to speak more directly to the original question of whether or not CRP payments negatively affected rural employment growth, *ceteris paribus*.

The second step entails choosing sample sizes over which the probability of a Type II error is calculated. The original study included 190 matched pairs. We include this sample size in the power analysis as a reference point. We also evaluate sample sizes 100 to 350 in steps of 50 observations. Varying the sample size and effect size results in a power surface indicating how the Type II error rate of the test, given sample size n and a posited effect size $\bar{\beta}_{CRP}$.

The third step of the *ex post* power analysis requires reconstructing the Data Generating Process (DGP) of the model. This step requires a simulation procedure similar to a bootstrap percentile t-test (Cameron and Trivedi, 2005). We apply a residual bootstrap procedure to simulate the DGP, which entails the following steps:

a) For sample size n and CRP effect size $\bar{\beta}_{CRP}$, resample with replacement from the original design matrix and OLS residual vector $(X_n, \hat{\varepsilon}_n)$ to generate a bootstrap replicate

data set $(X_n^*, \hat{\varepsilon}_n^*)$.

b) Calculate a new y_n as $y_n^* = X_n^* \hat{\beta} + \hat{\varepsilon}_n^*$, replacing $\hat{\beta}_{CRP}$ with $\bar{\beta}_{CRP}$.

c) Regress X_n^* on y_n^* with OLS.

d) Calculate the t-statistic (t^*) for $\hat{\beta}_{CRP}^*$ under the null, $H_0 : \beta_{CRP} = 0$.

e) For a 1-tailed test and a Type I error rate of $\alpha = 0.05$, if $t^* < t_{0.05,160}$ ($k = 30$, the number of covariates), then tally a rejection (r) of the null with a "1".

f) Return to step (a)

g) Repeat M times.

The power of the test is determined as r divided by the number of simulations in a sample size/effect size pair. We set $M = 10,000$ iterations for each combination, including the sample size of 190, which corresponds to original number of matched pair observations. The critical bound of the test statistic is $t_{0.05,160} = -1.654$, which corresponds with a cut-off of $\beta_{CRP(Crit)} = -1.654 \times 0.0034 = -0.0057$.

6 Ex Post Power Simulation from Posterior Marginal Distributions

An alternative but comparable approach for determining *ex post* power entails characterizing the entire posterior distribution of an estimated effect, along with the highest posterior density (HPD) bounds.[2] This Bayesian approach determines the posterior power of the 1-tailed test considered here by integrating underneath the HPD region left of the critical bound associated with the null hypothesis; i.e., -0.0057. The drawback of this approach is that the effect of sample size on power cannot be determined.

[2]We thank an anonymous reviewer suggesting this approach.

The normal regression model is

$$y_i \sim N(X_i\beta, \sigma^2) \tag{3}$$

$$\pi(\beta_k) \sim N(0, \sigma_k^2)(k \text{ does not include } \beta_{CRP}) \tag{4}$$

$$\pi(\beta_{CRP}) \sim N(\bar{\beta}_{CRP}, \sigma_{CRP}^2) \tag{5}$$

$$\pi(\sigma^2) \sim IG(a, b) \tag{6}$$

where the π's are prior distributions assigned to the model parameters, and IG is the inverse gamma distribution with shape and scale parameters (a, b). We use diffuse priors for the variance of the β_k's and σ^2, setting σ_k^2 to 10,000 and (a, b) to 0.01 (LeSage and Pace, 2009). Strong priors are used on the CRP effect, anchoring the distribution of β_{CRP} to the posited effect size and using the square of the OLS standard error as prior for the variance (table 3, $\sigma_{CRP}^2 = 0.0034^2$).

For each effect size evaluated, 1,000 Markov Chain Monte Carlo (MCMC) samples are generated with a thinning interval of 10 and a burn-in period of 5,000. In other words, the first 5,000 draws were discarded, after which every 10^{th} draw was retained. This method helps reduce autocorrelation in the chain but decreases the time to convergence. Chain convergence was assessed with the effective sample size associated with each parameter and the minimum, maximum, and average efficiency across parameters.

Posterior power is estimated with the result of a closed-interval test based on the marginal posterior distribution of β_{CRP}. The closed-interval test is

$$H_0 : \beta_{CRP} \in (-\infty, -0.0057] \tag{7}$$

The probability the null hypothesis is true is enumerated as the count of the simulated β_{CRP} posterior estimates meeting this condition. The resulting probability $P(H_0)$ is therefore the

posterior power of the test, or 1−Pr[*Type II Error*].

7 Results

Table 4 summarizes the power of the test using the Monte Carlo resampling method for each sample and effect size combination. The grey shading highlights the results for the original study's sample size ($N = 190$). Conventionally, a Type II error rate of 0.20 corresponds with a powerful test (= 0.80) (Cohen, 1988; Bayarri et al., 2016).[3] For the effect size of -1.5 percent, the power was 0.99. With an effect size of -1 percent, the power of the test was 0.88. The power of the test is diminished at lower effect sizes (-0.001 and -0.005), requiring suspension of judgment at those levels. Power increases as sample size and effect size increase. The power of the test converges to 1.0 after 150 observations at the effect size of -0.027. This result indicates that if the decline in employment growth from CRP was enough to offset the program's environmental benefits, we would have nearly 100 percent chance of detecting a negative effect of CRP on employment growth.

A visual illustration of this finding is in Figure 1 (top panel), which shows the empirical distributions of 10,000 draws using 190 observations and CRP effect sizes of -0.005 and -0.015. The solid vertical line corresponds with the critical value of the 1-tailed t-test corresponding with rejection of the null hypothesis (-0.0057). The overlap of the tails of the alternative distribution generated with an effect size of -0.005 and the null distribution are indicative of a relatively low test power (= 0.42). This leads to a rejection ratio of 0.42/0.05 = 8.4 (see Bayarri et al. 2016, for a discussion on size, power, and rejection ratios). In other words, in repeated samples, the odds are 8.4:1 that the observed parameter was generated from the alternative, enough perhaps to inspire a yawn of confidence. The safe course of action would be to suspend judgement. As the effect size increases (more negative in this case), the

[3]The intuition is provided by Bayarri et al. (2016), among others. At a Type I error rate of $\alpha = 0.05$ and a power of 0.80, a rejection would be 16 times more likely to occur under the alternative than under the null in repeated samples. However, there is no agreement what level of power constitutes a powerful test; 0.80 is conventional.

distribution of the simulated coefficient shifts further leftward, corresponding with a higher power test. At the effect size of -0.015 percent, the power of the test is 0.99, as indicated by the decrease in the overlap of the null and alterative distribution tails. In this case, an estimate falling to the left of the critical value of -0.0057 percent is $0.99/0.05 = 19.8$ times more likely to arise under the alternative distribution than the null at this posited effect size. The power to discern an effect increases, along with confidence in conclusion drawn from the test.

The posterior distributions of the effect size diverged from the priors (table 5). In Bayesian analyses, this occurs when there is insufficient information in the sample data. For example, according to the Bernstein-von Mises theorem, the posteriors and prior distributions are independent in large samples. In practice, the dependence between the choice of priors and model results is assessed by sensitivity analysis. We therefore evaluated the normal model of equations 3 to 6 over a range of priors from-0.048 to 0 in increments of 0.01 (all normal distributions with a variance of 0.0034^2). Although the posterior estimates of the effect sizes differed from their respective priors, the correspondence is strong and positive (Pearson's correlation = 0.99).[4] In the foregoing example, a 1-percent increase in the level of the effect size prior corresponds with a 0.62 percent increase in the posterior estimate of the effect size.

The posterior distributions of the effects sizes matching those generated by the resampling procedure were selected for comparison at the $N = 190$ sample of the original ERS study (table 5, figure 1). The power curves of both approaches are comparable. At the effect sizes of -0.001 and -0.005, the test's power to detect an effect is lower as determined by the Bayesian approach. The rejection ratio exceeds 16:1 for both approaches at and above an effect size of -0.01. Taken together, these results indicate that the original model had, under conventional levels, good ($\geq$ 80 percent) power for detecting negative employment effects from CRP of 1-percent or lower. At effect sizes lower than -1 percent, the prudent decision

[4]For all models, the effective sample size averaged 0.97 and the acceptance rate was 1. We conclude the MCMC chains converged.

maker might choose to suspend judgment.

The draw-back of the Bayesian approach suggested here is that *ex post* power cannot be compared over different sample sizes. In addition, depending on how sensitive posterior distributions are to priors, additional computations over a wide range of priors may be required to examine the *ex post* power of a test at a specified effect size. An advantage of the Bayesian approach is that it is relatively straightforward to extend to discrete or censored variables. Routines in SAS (SAS Institute, 2016), Stata (StataCorp, 2017) and R (McElearth, 2016) are well-documented with code, examples, and technical support.[5] The Monte Carlo resampling approach would require additional assumptions with respect to the DGP's of error terms.

8 Conclusion

The 2004 Report to Congress on the economic impacts of the Conservation Reserve Program provides a textbook case of how economists using conventional econometric testing protocol can inform policymakers regarding potential unintended adverse consequences of a policy. The challenge presented by the ERS study was that the econometric analysis of long-term employment effects of CRP culminated in a null finding. The Report correctly cautioned that the econometric verdict of "not statistically significant" could not be directly interpreted as "no adverse effect" because the statistical power of the test was unknown. The report did provide corroboratory evidence that rural counties with high CRP enrollments might be adapting to reduced agricultural production via increases in recreational spending. The preponderance of evidence supported the conclusion of "no adverse effect" even if a concise statistical verdict was unavailable.

Replication of the 2004 analysis confirms the 190 matched pair sample exhibited good power (0.88 to 0.92) for detecting a moderate adverse effect of CRP enrollment on employment growth of -1 percent. Applying new methods of statistical inference to the data and

[5]Data and code are available as supplementary materials online.

model used in the 2004 Report allows a more definitive conclusion: that the absence of statistical significance can reasonably be interpreted as "CRP has no substantive effect," since the probability of a Type II, or false negative error, is 0.12 percent if a reduction of -1 percent in employment growth is dispositive.

The potential for a crisp statistical verdict for null findings hints at a potential increase in the productivity of economists whose econometric work is used to inform decision-making. Because the continued move to evidence-based policy making will emphasize the normative importance of nonsignificant findings (Nussle and Orszag, 2014), economists armed only with NHST will be limited in their ability to address the "nonsignificant finding lacking error probability" conundrum. The costs associated with this could include the collection of extra-statistical information to reinforce equivocal statistical inference and the possibility of increased data collection costs if increased sample size is regarded as reliable insurance against uninformative studies. Proposed impact studies with seemingly limited samples may be rejected out of hand even if powerful statistical tests might be supported by the data. The alternative approaches presented as a proof of concept here is consistent with Practice 9 of the *Principles and Practices of a Federal Statistical Agency* (National Academies of Science, Engineering, and Medicine, 2017, p. 101) to "*keep abreast of and use modern statistical theory and sound statistical and computational practices in all technical work.*" Elevation from proof of concept to sound statistical practice will require critical assessment and further development of contemporary versions of the Neyman-Pearson protocol and/or use of Bayesian techniques by the community of economists in both academia and government.

In many circumstances, economists do not have the opportunity to conduct *ex ante* power analysis before research starts. The approaches we suggest can be used to determine *ex post* power for univariate analyses or multivariate regressions if the data generating process can be replicated and if the effect size of economic significance or policy relevance is stated. Given a range of posited effect sizes, our approach supplements an array of tools to inform decision making in the event of a null finding.

References

Abadie, Alberto, and Guido W. Imbens. 2011. "Bias-corrected Matching Estimators for Average Treatment Effects." *Journal of Business and Economic Statistics*, 29(1): 1–11.

Bayarri, M.J., Daniel J. Benjamin, James O. Berger, and Thomas M. Sellke. 2016. "Rejection Odds and Rejection Ratios: A Proposal for Statistical Practice in Testing Hypotheses." *Journal of Mathematical Psychology*, 72 90–103.

Cameron, Colin A., and Pravin K. Trivedi. 2005. *Microeconomics: Methods and Applications*. Cambridge: Cambridge University Press.

Christensen, Ronald. 2005. "Testing Fisher, Neyman, Pearson, and Bayes." *The American Statistician*, 59(2): 121–126.

Cohen, Jacob. 1988. *Statistical Power Analysis for the Behavioral Sciences*. New Jersey: Lawrence Erlbaum Associates, 2nd edition.

Diamond, Alexis, and Jasjeet S. Sekhon. 2013. "Genetic Matching for Estimating Causal Effects: A General Multivariate Matching Method for Achieving Balance in Observational Studies." *Review of Economics and Statistics*, 95(3): 932–945.

Ho, Daniel E., Kosuke Imai, Gary King, and Elizabeth A. Stuart. 2007. "Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference." *Political analysis*, 15(3): 199–236.

Hoover, Kevin D., and Mark V. Siegler. 2008. "Sound and fury: McCloskey and significance testing in economics." *Journal of Economic Methodology*, 15(1): 1–37.

LeSage, James P., and Robert Kelley Pace. 2009. *Introduction to Spatial Econometrics*. Boca Raton, FL: CRC Press.

McCloskey, Deirdre N., and Stephen T. Ziliak. 1996. "The Standard Error of Regression." *Journal of Economic Literature*, 34(1): 97–114.

McElearth, Richard. 2016. *Statistical Rethinking: A Bayesian Course in R and Stan.* Boca Raton, FL: CRC Press.

National Academies of Science, Engineering, and Medicine. 2017. *Principles and Practices for a Federal Statistical Agency, Sixth Edition.* Washington, D.C.: The National Academies Press.

Nickerson, Cynthia, Timothy Park, John Pender, Timothy R. Wojan, J. David Brown, Christine Heflin, Susan Helper, Cassandra Ingram, C.J. Krizan, Paul Marck, Samantha Schasberger, Kenneth P. Voytek, Jonathan Simonetta, and Giuseppe Gramigna. 2017. "Building Smarter Data for Evaluating Business Assistance Programs: Guide for Practitioners."Technical report, U.S. Small Business Administration, Washington, D.C.

Nussle, Jim, and Peter Orszag. 2014. *Moneyball for Government.* New York: Disruption Books.

SAS Institute. 2016. *SAS.* Cary, North Carolina, USA.

StataCorp. 2017. *Stata Statistical Data Analysis 15.1.* College Station, Texas 77845 USA.

Sullivan, Patrick, Daniel Hellerstein, LeRoy Hansen, Robert Johansson, Steven Koenig, Ruben N. Lubowski, William D. McBride, David A. McGranahan, Michael J. Roberts, Stephen J. Vogel, and Shawn Bucholz. 2004. "The Conservation Reserve Program: Economic Implications for Rural America." Agricutural Economic Report Number 834, U.S. Department of Agriculture, Economic Research Service, Washington, D.C.

Tweeten, Luther. 1983. "Hypotheses Testing in Economic Science." *American Journal of Agricultural Economics*, 65(3): 548–552.

Wasserstein, Ron. 2016. "American Statistical Association Releases Statement on Statistical Significance and P-Values." *ASA News*.

Wojan, Timothy R., Jason P. Brown, and Dayton M. Lambert. 2014. "What to Do about the "Cult of Statistical Significance"? A Renewable Fuel Application using the Neyman-Pearson Protocol." *Applied Economic Perspectives and Policy*, 36(4): 674–695.

Ziliak, Stephen T., and Deirdre N. McCloskey. 2004. "Size Matters: The Standard Error of Regressions in the American Economic Review." *Journal of Socio-Economics*, 33(5): 527–546.

Table 1: Mean values of Industrial, Labor Market, and Farm Structure Variables

Variable Description	Unit	High-CRP Counties	Matched Counties
Local economic characteristics:			
Agricultural employment, 1980	Percent	31.7	24.7**
Manufacturing employment, 1980	Percent	5.7	8.4**
Mining employment, 1980	Percent	2.2	2.3
Business services employment, 1980	Percent	3.9	4.2*
Recreation employment, 1980	Percent	4.1	4.5*
Special development dummy variables[1]:			
Prison county	0/1	1	0
Casino county	0/1	0	1.5
Meatpacking plant county	0/1	0.5	1
Labor market and location characteristics:			
Civilian employment, age 15-64, 1980	Percent	64.9	65.6
Working outside the county, 1980	Percent	10.9	12.9*
Median household income, 1979	$	12,620	12,936
Adjacent to a metropolitan area, 1983	0/1	15.9	22.6
Great Plains county	0/1	80	59.5**
Agricultural characteristics:			
Cropland/all land, 1982	Percent	46.7	45.1
Irrigated farmland, 1982	Percent	4.3	8.5**
Grain/total sales value, 1982	Percent	38.4	31.5**
Wheat/total sales, 1982	Percent	25.2	12.2**
Livestock/total sales, 1982	Percent	51.5	61.6**
Govt. payments/total income, 1981-83	Percent	6	2.6**
CRP enrollment/cropland, 1991-93	Percent	21.3	5.1**
CRP payments/income, 1991-93	Percent	6.7	0.8**
Farm sales/household income, 1980	Percent	1.9	1.4**
Farms w/ sales over $250,000 in 1982	Percent	5.3	5.8
Farms w/ sales under $20,000 in 1982	Percent	35.7	38.9*
Farmers working off-farm 200+ days, 1982	Percent	17.9	21.0**

Notes: * and ** indicate that the difference between high-CRP counties and their matched pairs is significantly greater than 0 at the 0.05 and 0.01 level, respectively. High CRP counties have an average CRP rental-payment-to-income ratio for 1991-93 exceeding 2.75 percent.[1] Statistics reported are the percent of observations coded as "1." *Source: Reproduced from Sullivan et al. 2004, p. 80.*

Table 2: Mean values of Employment Trends, Demographic and Amenity Variables

Variable Description	Unit	High-CRP Counties	Matched Counties
Post-CRP employment change:			
1985-1992 (short run)	Percent	-3.7	1.4**
1985-2000 (long run)	Percent	7.6	13.4**
Pre-CRP employment change			
1970-1982 employment	Percent	1.6	13.5**
1982-1985 employment	Percent	-1.7	0.3**
Demographic characteristics:			
Black population, 1980	Percent	0.6	0.4
Hispanic population, 1980	Percent	4.4	6.9
Native American population, 1980	Percent	3.3	1.9
Population under 18, 1980	Percent	29.8	29.3
Population over 62, 1980	Percent	19.3	19.7
Under 12 years of school, aged 25-44, 1980	Percent	17.2	16.5
College grads, aged 25-44, 1980	Percent	16.9	17.4
Population density, 1980	Percent	5	10**
Natural amenity characteristics:			
High mountains dummy variable1	0/1	5.6	10.8
Water/total area (x 10)	Log	-6.5	-6.2
Land in forest	Percent	3.7	8.5**
January days with sun (x 10)	Z-score	5.2	5.4
January temperature (x 10)	Z-score	-8.3	-6.1*
July humidity (x 10)	Z-score	9.7	7.1**
July temperature (x 10)	Z-score	-4.8	-5
Natural amenities scale (x 10)	Z-score	-7.2	-6.6

Notes: * and ** indicate that the difference between high-CRP counties and their matched pairs is significantly greater than 0 at the 0.05 and 0.01 level, respectively. High CRP counties have an average CRP rental-payment-to-income ratio for 1991-93 exceeding 2.75 percent.[1] Statistics reported are the percent of observations coded as "1." *Source: Reproduced from Sullivan et al. 2004, p. 79.*

Table 3: Replication of Long-Run Job Growth Model

Variable	Beta	Std. Error	t-stat	Pr($>\|t\|$)	Standardized Beta[a]
CRP payments to income ratio	0.007	0.003	1.945	0.054	0.237
Population density, 1980	0.035	0.034	1.052	0.294	0.181
Density x CRP ratio	-0.002	0.003	-0.576	0.566	-0.061
Employed in ag, 1980	-0.002	0.002	-1.114	0.267	-0.159
Density x Percent ag emp.	0.000	0.001	-0.367	0.714	-0.046
Population, 1982/1970	0.256	0.195	1.314	0.191	0.158
Population, 1985/1982	0.225	0.314	0.716	0.475	0.056
Employment, 1982/1970	-0.175	0.086	-2.039	0.043	-0.204
Employment, 1985/1982	-0.157	0.163	-0.964	0.337	-0.075
Under 18 years of age, 1980 (%)	0.006	0.006	1.039	0.300	0.157
Over 62 years of age, 1980 (%)	0.000	0.005	-0.036	0.971	-0.005
American Indian, 1980 (%)	0.002	0.002	1.097	0.274	0.115
Black, 1980 (%)	-0.008	0.002	-3.220	0.002	-0.231
Hispanic, 1980 (%)	0.001	0.002	0.700	0.485	0.079
Cropland, 1982 (%)	-0.001	0.001	-1.270	0.206	-0.155
Livestock/total sales, 1982	0.000	0.001	-0.694	0.489	-0.063
Govt payments/income, 1981-83	-0.005	0.005	-1.051	0.295	-0.131
Wheat/total sales, 1982	-0.001	0.001	-1.119	0.265	-0.117
Less than high school, 1980	-0.002	0.002	-0.958	0.339	-0.123
College, 1980	0.002	0.003	0.550	0.583	0.046
Civilian employment rate, 1980	0.002	0.003	0.760	0.449	0.069
Median household income, 1979	-0.070	0.097	-0.723	0.471	-0.080
Natural amenities index	-0.004	0.013	-0.321	0.749	-0.027
Land in forest (%)	0.002	0.001	2.389	0.018	0.253
Great Plains county (1/0)	-0.036	0.028	-1.264	0.208	-0.116
Employed in mining, 1980 (%)	-0.027	0.026	-1.062	0.290	-0.075
Employed in recreation, 1980 (%)	0.003	0.007	0.449	0.654	0.035
Commuting outside county, 1980	0.001	0.002	0.725	0.469	0.058
Meat packing plant county (1/0)	0.042	0.092	0.460	0.646	0.030
Casino county (1/0)	0.139	0.123	1.136	0.258	0.070
Prison county (1/0)	-0.019	0.083	-0.223	0.824	-0.015
N	190				
Adj. R^2	0.341				
F-stat	4.177	p-value	0.000		

Note: [a]The last column of the table is for comparison to the standardized coefficients reported in Table A.3 on p. 82 of Sullivan et al. (2004).

Table 4: Simulated Power of One-Tailed Test by Sample and Effect Size

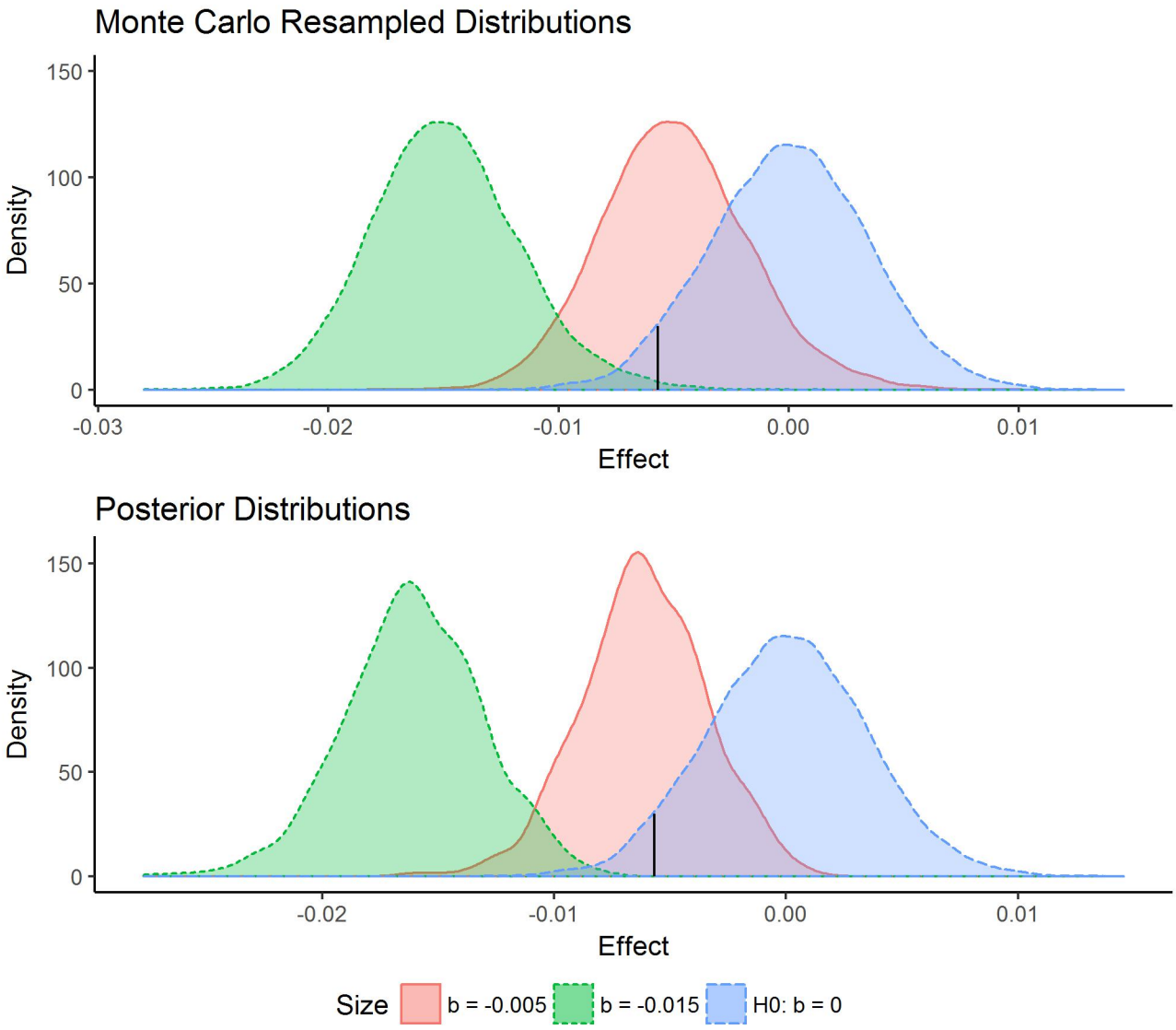
	Sample Size[a]						
Beta for CRP	100	150	190	200	250	300	350
-0.001	0.06	0.06	0.06	0.06	0.07	0.07	0.07
-0.005	0.24	0.33	0.42	0.43	0.51	0.60	0.67
-0.01	0.59	0.79	0.88	0.90	0.96	0.98	0.99
-0.015	0.84	0.96	0.99	0.99	1.00	1.00	1.00
-0.027	0.99	1.00	1.00	1.00	1.00	1.00	1.00

Notes: The grey shading corresponds to the sample size in Sullivan et al. (2004).
[a]Power was calculated from 10,000 draws of each sample size and re-estimation of the model using the MCMC resampling procedure.

Table 5: Simulated Posterior Power of One-Tailed Test and Effect Size

Prior	Posterior estimate of Beta CRP	Lower HPD[b]	Upper HPD	Posterior power[b]
-0.009	-0.001	-0.006	0.004	0.04[c]
-0.016	-0.005	-0.010	0.000	0.39
-0.024	-0.010	-0.015	-0.004	0.92
-0.032	-0.015	-0.021	-0.009	1.00
-0.048	-0.027	-0.034	-0.021	1.00

Notes: [a]Highest posterior density (HPD) 95% credible interval.
[b]Gibbs MCMC power stimates evaluated at N = 190 (MCMC sample size = 1,000, with a thinning break of 10).
[c]Estimated power is less than theoretical minimum power due to random error. Error bounds are calculated as $\pm 2 \times \sqrt{(0.05 \times 0.95/1,000)} = [0.0364, 0.064]$.



Note: The null distributions are identical in each panel, simulated as $\sim N(0, \sigma^2)$, with the variance of the OLS standard error estimate of the CRP variable in Table 3. The vertical black line is the critical value $-1.654 \times 0.0034 = -0.0057$ at the 5% level for a 1-tailed t-test (159 degrees of freedom) under the null. The alternative distributions in the top panel are from the Monte Carlo resampling procedure (10,000 replicates). The alternative distributions in the bottom panel are posterior marginal distributions of the CRP estimate (1,000 samples, with a thinning break of 10).

Figure 1: Empirical distribution of estimates from imposed effect sizes at $N = 190$