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

Jason P. Brown
Federal Reserve Bank of Kansas City
Jason.Brown@kc.frb.org

Dayton M. Lambert
University of Tennessee Institute of Agriculture
Department of Agricultural & Resource Economics
Knoxville, TN
dlamber1@utk.edu

Timothy R. Wojan
Economic Research Service/USDA
twojan@ers.usda.gov

Selected Paper prepared for presentation at the Agricultural & Applied Economics Association's 2017 AAEA & CAES Joint Annual Meeting, Chicago, IL, July 30-August 1, 2017.

[*] The opinions expressed are those of the authors and are not attributable to the Federal Reserve Bank of Kansas City, the Federal Reserve System, the University of Tennessee, the Economic Research Service, or the U.S. Department of Agriculture. The authors wish to thank Dan Hellerstein, David McGranahan, and Pat Sullivan for discussions about the earlier ERS research replicated in this paper. All errors are the sole responsibility of the authors. Lambert's research effort was partially funded by the USDA Hatch Multi State Project, NE-1049, "Community Health and Resilience".

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

Abstract

The objective of this paper is to provide error probabilities for null findings, allowing applied economists to more confidently conclude when "not significant" can in fact be interpreted as "no substantive effect." The example used to demonstrate our method is the Economic Research Service's (ERS) 2004 Report to Congress that was charged with statistically identifying any unintended negative employment consequences of the Conservation Reserve Program (CRP) (Sullivan et al., 2004). The report failed to identify a statistically significant negative effect of the CRP on employment growth, but the authors correctly cautioned that the verdict of "no negative employment effect" was only valid if their 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 the 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 econometric test used in the report had very high power for detecting large negative employment effects from CRP and sufficiently high power for detecting a modest effect. Paradoxically, the unrestricted charge to search for "any effect" had very low power.

JEL Codes: C12, Q42, R11

Key words: power analysis, Monte Carlo simulation, hypothesis testing

Introduction

The objective of this paper is to provide error probabilities for null findings, allowing applied economists to more confidently conclude when "not significant" can in fact be interpreted as "no substantive effect." The example used to demonstrate our 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 declines in many farm-dependent counties through the 1990s raised concerns 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 that characterize the great 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 estimate 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 small, moderate, and large adverse employment effects

following enrollment of cropland into the CRP. The second deficiency is conceptually and computationally more challenging. We develop candidate distributions by 'baking' an effect size into simulated data whose error structure is recovered from sample data.

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 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, then our replication reinforces the original findings. Since the test to detect a moderate effect was powerful, the null finding can be interpreted as "no economically significant effect." The broader implications for econometrics practice are discussed in the conclusion.

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 and 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 where compelling alternative hypotheses are unknown. The benefits this approach provides are immediate: 1) only a single distribution is required for testing whether 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 approach is that no statistical inference is possible for a nonsignificant result: the proof by contradiction has failed and so the only valid verdict is to suspend judgment. But this cost in scientific exploration in its purest sense 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 the breakeven point for a treatment or policy. The research must also posit what the alternative distribution looks like, traditionally provided by a literature search of studies of similar phenomena. 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 degrees of confidence, the treatment effect is less than* <posited effect size>."

The Neyman-Pearson protocol is perhaps foreign to many economists, but this approach won out in applied statistical disciplines where equivocal findings could impose significant monetary costs, such as biomedical research. The most persuasive explanation for why econometrics dismissed Neyman-Pearson in favor of Fisherian NHST is the much greater difficulty economists have speculating about an unobserved alternative distribution (Wojan,

Brown, and Lambert 2014). Lacking conjectures on 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, applied econometrics grafted the concepts of statistical power and Type II error (falsely failing to reject the null hypothesis) onto the Fisherian NHST protocol as issues of concern. But importantly, these concerns were not things that would ever be estimated.

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. A more sophisticated approach is something that could be called the heuristic of equivalent power. If an analysis of a given sample size produces statistically significant results, then tests for all other specifications and dependent variables in the analysis are deemed adequately powerful. 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 quite 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 has 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, Brown and Lambert 2014). Transparency in the 2004 CRP Report included the caveat that the statistical power of the econometric test was not known (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 3 of the *Statement of Commitment to Scientific Integrity* by Principal Statistical Agencies (2012) that addresses "openness about sources and limitations of the data provided." The question is whether the assumed infeasibility of statistical power assessments is truly binding. To answer this question, we provide a concrete example of generating *ex post* statistical power estimates using new methods to aid the interpretation of nonsignificant findings regarding a public policy issue.

Congressionally Mandated Study of Economic Implications of the Conservation Reserve Program

The Conservation Reserve Program was first authorized in 1985 for the purpose of providing public benefits by taking environmental 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, there may be less concern that the program would have any 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 that the decline in agricultural production would not be fully compensated by CRP payments in the local economy. And since many counties with a large number of 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 ERS study provides a comprehensive examination of the effect of CRP on farm and non-farm aspects of the rural economy 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. The comprehensiveness of the report helps reinforce the story that the statistical analysis of employment trends supported but could not definitively confirm: i.e., that implementation of 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 we hope to resolve in the present study.

Testing for the effects of high-CRP enrollment on employment growth presented the daunting 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, 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 the difference 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 the difference in matching variables that persisted to isolate the effect of high versus low CRP enrollments on job growth.

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 been able to find much closer matches on all these variables, then simply comparing the average employment growth across groups would be informative of the impact of high-CRP enrollments. However, given the difference in structural characteristics it is reasonable to assume that many factors other than

CRP enrollment contributed to differences in job gains. Table 2 demonstrates that these differences were large—an employment change difference of 5.8% between high-CRP and matched counties.

<< Table 1 >>

<< Table 2 >>

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-run regressions did find negative impacts of high-CRP on employment growth that were statistically significant in 7 out of 20 alternative specifications (magnitudes of these estimates were not provided). Applying the heuristic of equivalent power, this was suggestive that the relatively small sample size of 190 matched pairs was adequately powerful. However, the heuristic was not invoked in the report—its main purpose 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 different specifications estimated for the long-run 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 3) outlined in the *Statement of Commitment to Scientific Integrity by Principal Statistical Agencies:*

> *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 equivalent power was replaced by the more econometrically justifiable heuristic of robustness. While robustness tests often provide valid checks of empirical findings, when applied to questions of statistical power they devolve to mere heuristics. If a statistical test is in fact weak, numerous re-specifications will only provide additional evidence of weak power. Numerous re-specifications do not reinforce the erroneous conclusion that not significant can be interpreted as "has no effect."

Deriving a Statistical Answer

Given the statement from the report above, 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 nonsignificant 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 over the last 20 years. McCloskey and Ziliak (1996, p. 105) examined the issue quantitatively and found that fewer than 30 percent of papers 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 papers that had "consider[ed] the power of the test" may have done this. By the 1990s, 8 percent of papers published in *AER* considered the power of the test suggesting a very modest improvement among elite economists (Ziliak and McCloskey 2004).

Congress charged ERS with identifying any negative impacts. Positing an effect size for the purpose of analysis could be interpreted as inconsistent with Congressional intent as the effect size that mattered was explicit: *any effect*. 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 as the final decision regarding what matters is retained by the policymaker.

Describing a worst case scenario for unintended adverse effects of the program provides a compelling case of what would constitute a large effect size. Job losses equivalent to the

environmental benefits of the program qualifies as such a scenario. Arriving at an estimate for this figure is all that is required as 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) we assume pure controls (no CRP acres in low-CRP counties); and 4) one job in 2000 in the treatment counties can be valued at $23,897—average earnings per nonmetro job derived from the Bureau of Economic Analysis. Arriving at a ballpark employment loss percentage of a large effect is calculated simply as 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):

Job Equivalent Cost of Program × Treatment Counties Share of Program × [1/Treatment County Jobs] =

$$\frac{\$23.7 \text{ Billion}}{\$23{,}897}\ (a) \times \frac{508{,}000 \text{ acres}}{33{,}981{,}000 \text{ acres}}\ (b) \times \frac{1}{537{,}398}\ (c) = 2.76\%$$

These grossly simplified—though reasonable—assumptions provide useful information for characterizing a large adverse effect of the program. Remembering that employment growth in treatment counties lagged that of 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, more moderate adverse effects could also provide an economic basis for modifying CRP. Effect sizes roughly a half,

third, or fifth of the worst case scenario would correspond to effect size of 1.5 percent, 1 percent and 0.5 percent, respectively. To assess the power of detecting "any effect" an effect size of 0.1 percent—a degree of magnitude less than what could be considered a moderate effect—is also included in the estimates of power.

Arriving at a credible estimate of an unobserved alternative distribution is technically much 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 statistical power analysis: 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 traditional approach was impossible. The lack of prior studies led instead to a Monte Carlo simulation approach where the observed sample provides information needed to create the alternative distribution of interest (Sullivan, et al. 2004; Wojan, Brown and Lambert 2014).

In the current analysis, the purpose of a Monte Carlo analysis is to model the data generating process (DGP) of the outcome variable *y*, that is, the difference in county-level employment growth between treatment and control counties. As a first step in the simulation we begin with replicating the estimation of the long-run local employment growth model used to evaluate CRP in Sullivan et al. (2004). Although the authors reported the CRP estimates for several models, only complete results were reported for one version. We selected that specification so that our base results used to construct the Monte Carlo simulation would be as close as possible to the original model specification. Employment growth between 1985 and 2000 was estimated using ordinary least squares (OLS) on the differenced values between matched pairs of high-CRP and low-CRP counties with the linear model:

$$y_i = \left[ln\left(\frac{emp_{i,2000}^{HCRP}}{emp_{i,1985}^{HCRP}}\right) - ln\left(\frac{emp_{i,2000}^{LCRP}}{emp_{i,1985}^{LCRP}}\right)\right] = X_i\beta + \varepsilon_i, \quad (1)$$

where *i* indexes the matched pair, $X_i = \left(X_i^{HCRP} - X_i^{LCRP}\right)$ is a vector containing differences information on CRP payments and conditioning measures including local socioeconomic and agricultural characteristics shown in Tables 1 and 2, and ε is independent and identically distributed random component with mean zero and a constant variance. Equation 1 was estimated with the 190 matched pairs from the original study. Replication of the OLS estimation is reported in Table 3. Our results were nearly identical to those reported in Table A.3 of Sullivan et al. (2004). 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.

<< Table 3 >>

Using the estimation results, the next step in developing the simulation is to reconstruct the data generating process (DGP) of the model. The DGP is simulated by specifying: 1) a sample size; 2) the distribution of *X*, and ε; and 3) the values of β. Rather than commit to underlying distributions for *X* and ε, we instead bootstrap the observed values of *X* from the original matched pairs in the data and ε from the estimation of (1). The starting values of β also come from OLS estimation of Eq. 1, with the exception of the coefficient on CRP, which takes on the value of alternative hypotheses we are trying to detect. For each bootstrap replicate, we reconstruct the dependent variable as:

$$y_i^* = X_i^* \beta + \varepsilon_i^*, \quad (2)$$

where the X_i^* and ε_i^* are bootstrapped pairs sampled with replacement from the original data and OLS residuals of (1).

Repeated draws from the DGP are used to calculate the number of times (*r*) the null hypothesis for the coefficient of interest (CRP in this case) is rejected, given a Type I error rate of α. The power of the test is determined by *r* divided by the number of simulations of the DGP. We use sample sizes of 100 to 350 in steps of 50 observations with 10,000 iterations. In addition, we include a sample size of 190, which corresponds to original number of matched pair observations. The critical value α was set to the 5 percent level of significance for a one-tailed *t*-tests on the CRP coefficient, HA: $\beta_\delta < 0$, where δ is a posited effect size.

Selecting the value of β for CRP in the Monte Carlo simulation requires selecting alternative hypothesis for detection in the model estimation. Generally, effect size should correspond to the smallest economic effect that would matter. In this example a range of effect sizes is used 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 β_δ = -0.001, -0.005, -0.010, -0.015, and -0.027. These would correspond to 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 what effect sizes to test for 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 benefit and smaller effect sizes approaching zero. We specifically chose negative

numbers in order to speak more directly to the original question of whether or not CRP payments negatively affected rural employment growth.

For each sample size and CRP effect size combination, we calculate the power of one-tailed *t*-tests to detect that effect size in 10,000 simulations. Figure 1 shows the estimated power curves for each combination. As expected, power increases in sample size and effect size. However, at the smallest effect size (-0.001) the power converges close to the Type I error rate as the sample size increases. Looking at the power curve for the largest effect (-0.027), the power of the test converges to 1.0 after 150 observations. This result indicates that if the decline in employment growth from CRP was large enough to offset the benefits, we would have nearly 100 percent chance of detecting a negative effect of CRP on employment growth.

<< Figure 1 >>

A closer look of the power estimates is provided in Table 4. The grey shading highlights the results for a sample of 190 observations; the original sample size. With an effect size of -1 percent, the power of the test was 0.82. For 50 percent larger effect size of -1.5 percent, the power was 98 percent. At -0.05 percent and smaller, the power for finding an effect was relatively week. A visual illustration of this in shown in Figure 2, which shows the empirical distributions of 10,000 draws of the model using 190 observations and CRP coefficients of -0.001 and -0.01. The dashed and solid vertical lines correspond to the tenth percentile of each distribution of coefficients, which is approximate to a one-tailed *t*-test. It is clear to see that there is large overlap in the distribution of coefficients. As the effect size gets larger (more negative in this case), the distribution of simulated coefficients would shift further and further to the left

corresponding to a higher power test. Taken together, our results indicate that the original model had very high power for detecting large negative employment effects from CRP, sufficiently high power for detecting a modest effect, and very low power for finding a slightly greater or less than zero effect.

<< Table 4 >>

<< Figure 2 >>

Conclusion

The 2004 Report to Congress on the economic impacts of the Conservation Reserve Program provides a textbook case of how economists using traditional methods can inform policymakers regarding potential unintended adverse consequences of a policy. The challenge presented by the 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.

Our replication of the 2004 analysis confirms the 190 matched pair sample had adequate power for detecting a moderate adverse effect of CRP enrollment on employment growth of -1.5 percent. Applying new methods of statistical inference to data used in the 2004 Report allow 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 2 percent if a reduction of -1.5 percent in employment growth is dispositive.

The potential for a crisp statistical verdict for null findings hints at potentially large increases in the productivity of applied 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 (Ayotte, et al. 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 evaluations of programs with seemingly limited samples may be rejected out of hand even if powerful statistical tests might be supported by the data. The alternative presented as a proof of concept here is consistent with Practice 7 of the *Statement of a Commitment to Scientific Integrity* to *"keep abreast of and use modern statistical theory and sound statistical practice 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 by the community of economists in both academia and government.

References

Ayotte, K.; Warner, M.; Hubbard, G.; Sperling, G.; and Barnes, M. (2014). Moneyball for Government. Results for America, Disruption Books.

Christensen, R. (2005). Testing Fisher, Neyman, Pearson and Bayes. The American Statistician, 59, 121-126.

Hoover, K.D. and Siegler, M.V. (2008) Sound and fury: McCloskey and significance testing in economics. *Journal of Economic Methodology* 15(1):1-37.

Hymans, S. (2009). Review of The cult of statistical significance: How the standard error costs us jobs, justice and lives by Stephen T. Zilak & Deirdre McCloskey. Journal of Economic Literature, 47, 499-503.

McCloskey, D.N. & Ziliak, S.T. (1996) The standard error of regressions, *Journal of Economic Literature* 34(1): 97-114.

Nickerson, C., Park, T., Pender, J., Wojan, T., Brown, J.D., Heflin, C., Helper, S., Ingram, C., Krizan, C.J., Marck, P., Schasberger, S., Voytek, K.P., Simonetta, J., Gramigna, G. (2017). *Building Smarter Data for Evaluating Business Assistance Programs: Guide for Practitioners*, Washington, DC: U.S. Small Business Administration (May). Available at https://www.sba.gov/sites/default/files/aboutsbaarticle/Building_Smarter_Data1.pdf

Principal Federal Statistical Agencies (2012) "Statement of Commitment to Scientific Integrity," Administrative Publication No. (AP-059). Washington, DC: Economic Research Service. Available at https://www.ers.usda.gov/webdocs/publications/42787/31977_ap059.pdf?v=41171.

Sullivan, P., Hellerstein, D., Hansen, L., Johansson, R., Koenig, S., Lubowski, R., McBride, W., McGranahan, D., Roberts, M., Vogel, S., and Bucholtz, S. (2004) *The Conservation Reserve Program: Economic Implications for Rural America.* Agricultural Economic Report Number 834. Washington, DC: U.S. Department of Agriculture, Economic Research Service.

Tweeten, L. 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*, March 7. Accessed 5/23/17 at https://www.amstat.org/asa/files/pdfs/P-ValueStatement.pdf

Wojan, T.R., Brown, J.P & Lambert, D.M. (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, S.T. & McCloskey, D.N. (2008). The cult of statistical significance: How the standard error costs us jobs, justice and lives. Ann Arbor: The University of Michigan Press.

Ziliak, S.T. & McCloskey, D.N. (2004). Size matters: The standard error of regressions in the American Economic Review. Journal of Socio-Economics, 33, 527-46.

Tables

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 employment2	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 variable[1]	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[1]
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: [1]The last column of the table is for comparison to the standardized coefficients reported in Table A.3 on pg. 82 of Sullivan et al. (2004).

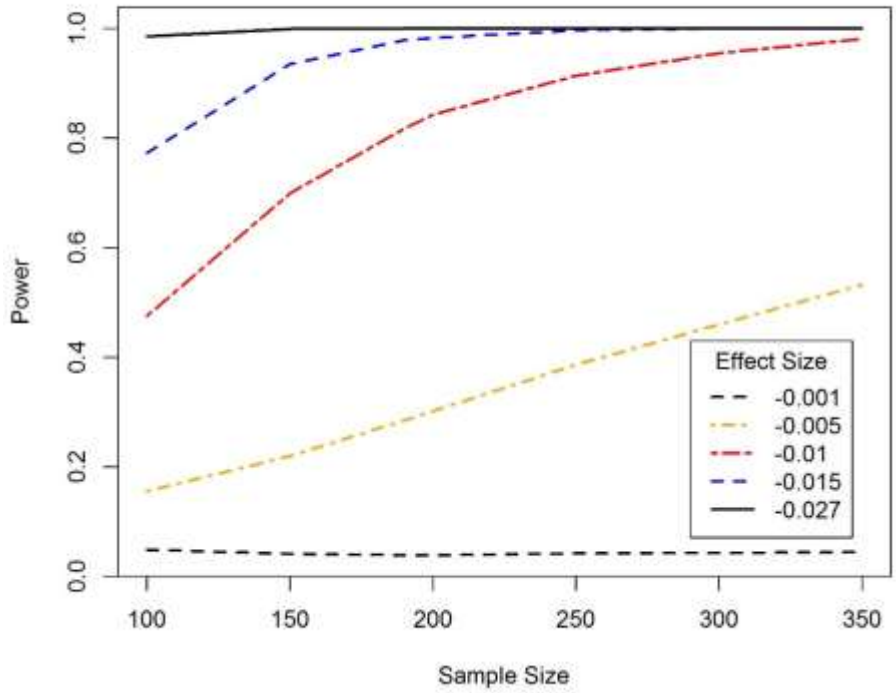
Table 4. Simulated Power of One-Tailed Test by Sample and Effect Size

	Sample Size						
Beta	100	150	190	200	250	300	350
-0.001	0.05	0.04	0.04	0.04	0.04	0.04	0.04
-0.005	0.16	0.22	0.28	0.30	0.39	0.46	0.53
-0.010	0.48	0.70	0.82	0.84	0.91	0.95	0.98
-0.015	0.77	0.94	0.98	0.98	1.00	1.00	1.00
-0.027	0.98	1.00	1.00	1.00	1.00	1.00	1.00

Note: The power was calculated from 10,000 draws of each sample size and re-estimation of the model. The grey shading corresponds to the sample size in Sullivan et al. (2004).

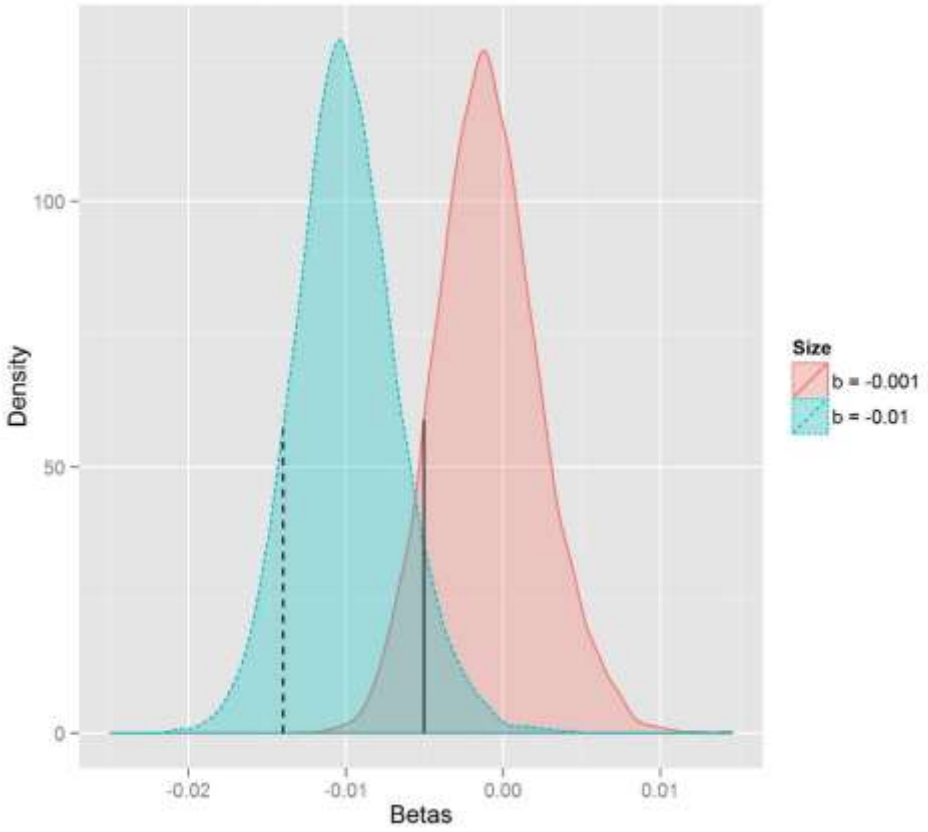
Figures

Figure 1. Simulated Power Curves



Note: Power was calculated from 10,000 simulations of the model for each sample and effect size combination.

Figure 2. Empirical Distribution of Estimates from Imposed Effect Sizes



Note: Each distribution shows 10,000 simulations of the model using 190 observations and the CRP coefficients.