Are All Economic Hypotheses False?

Journal of Political Economy, vol. 100, n° 6, Centennial Issue, 1992

J. Bradford De Long

Harvard University and National Bureau of Economic Research

Kevin Lang

Boston University and National Bureau of Economic Research

We develop an estimator that allows us to calculate an upper bound to the fraction of *unrejected* null hypotheses tested in economics journal articles that are in fact true. Our point estimate is that none of the unrejected nulls in our sample is true. We reject the hypothesis that more than one-third are true. We consider three explanations for this finding: that all null hypotheses are mere approximations, that data-mining biases reported standard errors downward, and that journals tend to publish papers that fail to reject their null hypotheses only when the null hypotheses are likely to be false. While all these explanations are important, the last seems best able to explain our findings.

I. Introduction

With the exception of a small minority of courses taught by Bayesian statisticians, most courses on econometrics and statistics teach classical hypothesis testing. In classical hypothesis testing, a null hypothesis is posed against an alternative, and the null hypothesis is considered "rejected" or "not rejected" on the basis of whether a single test statistic exceeds some critical value (e.g., whether a large-sample t-statistic

We thank Gary Koop, Edward Leamer, Kevin Murphy, the late George Stigler, Jim Stock, and an anonymous referee for helpful comments. This research was supported in part by the NBER and the Olin Foundation (De Long) and by a Sloan Foundation Faculty Research Fellowship (Lang).

[*Journal of Political Economy*, 1992, vol. 100, no. 6]
 0022-3808/92/0006-0002$01.50

exceeds 1.96). Although the falsificationist view of scientific methodology (Popper 1959) stresses the importance of specifying test events that *cannot* occur under the null hypothesis, economists are typically able to specify only events that are *unlikely* under the null hypothesis. Therefore, if the null is "rejected," our confidence in it is reduced; if the null hypothesis "fails to be rejected," our confidence in the correctness of the null hypothesis is increased because the data do not speak strongly against it.[1]

Since no individual test is definitive, it is somewhat surprising that the rhetoric of article writing suggests that a single test or series of tests in the individual article is conclusive. Clearly there is a role for papers that reconsider the corpus of empirical work on a topic. To a certain extent, this is the role of well-crafted review papers.

However, even a careful review of the existing published literature will not provide an accurate overview of the body of research in an area if the literature itself reflects selection bias. To take an extreme example, suppose that journals are interested only in publishing and authors circulate papers only with statistically significant results. An unbiased reviewer analyzing tests of a particular null hypothesis would observe that all the *published* tests had rejected the null. The reviewer would infer that the evidence against the null was quite strong. Such a conclusion might be very misleading if there were a substantial number of hypothesis tests that had in fact failed to reject but that were not circulated.

This form of publication bias, known as the "file drawer problem," has received considerable attention outside of economics in the literature on meta-analysis. For example, Berlin, Begg, and Louis (1989) note that if publication bias favors significant results, those published papers that use smaller sample sizes will tend to find larger effects, for larger effects are required for significance when the sample size is small. They confirm this relation for a sample of clinical cancer trials.

In this paper, we develop an approach that allows us to measure the fraction of *unrejected* null hypotheses that are, in fact, *false*. Our point estimate is that *none* of the unrejected nulls in our sample is true. Moreover, we can reject at the .05 level the null hypothesis that more than about one-third of *unrejected* null hypotheses are true.

We consider three prime contenders as explanations for this finding: (1) "So what? We know that all hypotheses are false; they are

[1] The distinction between the "standard" approach to testing in "science" and that used in economics should not be exaggerated. Many other sciences explicitly use statistical methods. The inferences drawn by researchers who do not are still influenced by their confidence in auxiliary hypotheses.

only approximations." (2) "So what? We all know that data mining negates the classical distribution of the t-statistic, so why be surprised that test statistics fail to conform to the distribution classical theory claims they should have?" We are sympathetic to both these views, and both have significant implications for the way economists do and evaluate empirical research.

However, we argue that a file drawer problem somewhat analogous to that discussed in the meta-analysis literature provides the most important explanation for our findings. (3) t-statistics and similar test statistics with values near zero are unlikely to be reported in the literature. This nonreporting could arise because authors do not report, and referees and editors do not demand, formal test statistics when the data speak in obvious ways. But we do not think that this is the principal cause of the phenomenon.

On the other hand, if authors and journals are more likely to publish statistically significant results, we can ask under what circumstances papers that do not report statistically significant rejections of the null hypothesis would make it past the publication filter. In medicine, recent findings that fetal monitoring is not an effective technology would have been much less interesting were it not widely believed that fetal monitoring is effective (in part because of previous studies that "establish" its effectiveness). This medical analogy suggests that studies that fail to reject their null hypotheses are much more likely to be published when prior work has already strongly established a contrary result. This makes it plausible that papers that fail to reject their null hypotheses survive the refereeing process and get published only if the probability that the null hypotheses they test are false is high, for when the null hypothesis is in fact false, earlier work is most likely to have established the contrary presumption that makes the paper's failure to reject interesting.

Section II outlines our basic approach to determining the fraction of unrejected null hypotheses that are, in fact, false. Section III presents our data. The principal findings are described in Section IV. Section V assesses the three potential explanations for our findings. Section VI summarizes our conclusions and poses the peculiar dilemma our findings pose for lines of empirical research that rely on failures to reject null hypotheses as confirmatory evidence.

II. Our Approach

Most empirical work in economics tests a null hypothesis against an alternative by calculating the marginal significance level associated with some test statistic. If the marginal significance level is less than some prespecified level (typically .05), then we say that we reject the

null hypothesis H_0 and conclude provisionally that it is false, in favor of the alternative hypothesis H_1, which we provisionally conclude is true. If the marginal significance level exceeds the critical value, then we say that we have "failed to reject" the null hypothesis H_0 and our confidence in H_0 is increased.

As we all learned in our first statistics class, such a decision procedure is subject to two types of errors. First, we can erroneously reject a true null hypothesis H_0 because an unlikely realization of the underlying random process has led to a low marginal significance level. It has become customary to set the critical marginal significance level at .05, so that when the null hypothesis H_0 is true such type I errors occur only 5 percent of the time: the size of the test is 5 percent.

Second, we can erroneously fail to reject a false null hypothesis H_0 when the alternative H_1 is in fact true. As a rule the critical significance level is not adjusted for the probability of such type II errors. Typically, the alternative hypothesis H_1 is diffuse and the test statistic has a different distribution for each point in H_1. Calculating the distribution under the alternative in order to construct hypothesis tests of a specified high power—and low chance of a type II error—requires, it is argued, more knowledge of the distribution of the test statistic under H_1 than the data can provide.

Economists' statistical tests, therefore, typically have a known size of 5 percent but an unknown power q. There is a tight bound on the chance of a type I error. If the null hypothesis H_0 is true in a fraction π of hypothesis tests, then the fraction of hypothesis tests that produce a type I error—land in the upper right box of figure 1—is $.05\pi$, which must be less than or equal to .05. By contrast, there is no analogous tight bound on the chance of a type II error—of failing to reject a false null hypothesis H_0 and landing in the lower left box of figure 1.

In this paper, we examine a large number of hypothesis tests that have been carried out in the past few years in order to learn about the fraction π of null hypotheses H_0 that are true and about the

	Fail to Reject H_0	Reject H_0	
H_0 True	$.95\pi$	$.05\pi$	π
H_1 True	$(1 - q)(1 - \pi)$	$q(1 - \pi)$	$1 - \pi$
	$1 - q + (q - .05)\pi$	$q + (.05 - q)\pi$	

FIG. 1.—Possible outcomes

average power, q, of economists' hypothesis tests. We conclude that π is essentially zero: that only a very small fraction of the null hypotheses in published articles are true. Failures to reject nulls are therefore almost always due to lack of power in the test, and not to the truth of the null hypothesis tested.

Our method relies on the fact that under the null hypothesis H_0, the marginal significance level of a test statistic is uniformly distributed over [0, 1] and satisfies

$$P(f(a) \geq p) = 1 - p, \tag{1}$$

where f is the marginal significance level of the calculated test statistic a. In other words, under the null we should fail to reject at the .9 level 10 percent of the time, fail to reject at the .8 level 20 percent of the time, and so on.

Under the alternative H_1, $f(a)$ has some unknown cumulative distribution

$$P(f(a) \geq p) = 1 - G(p). \tag{2}$$

We assume that the density $g(p)$ under the alternative is decreasing in p in such a manner that $[1 - G(p)]/(1 - p)$ falls monotonically from one at $p = 0$ to $g(1)$ at $p = 1$. Thus the chances of obtaining a value of $f(a)$ below any particular critical level and rejecting the null are greater under the alternative than under the null. The more stringent the required critical level, the greater the proportional differential between the probability of rejecting the null when it is true and the probability of rejecting the null when it is false.

We use (1) and (2) to write the unconditional distribution of $f(a)$ in terms of the distribution G and the unknown fraction π of null hypotheses H_0 that are in fact true:

$$P(f(a) \geq p) = \pi(1 - p) + (1 - \pi)[1 - G(p)]. \tag{3}$$

Since the cumulative distribution $G(p) \leq 1$ for all p in [0, 1],

$$\pi \leq \frac{P(f(a) \geq p)}{1 - p}. \tag{4}$$

Equation (4) allows the construction of an upper bound on the fraction π of null hypotheses that are true. For every critical value p, the fraction of reported test statistics with marginal significance levels at or above p provides us with an estimate of the numerator of (4); if one-half of all null hypotheses tested are true, then at least one-tenth of marginal significance levels $f(a)$ ought to be above .8.

The bound (4) is tightest for values of p near one, for there the density $g(p)$ under the alternative is lowest. The bound (4) becomes

trivial for values of p near zero: at $p = 0$, equation (4) becomes $\pi \leq 1/1$, which is always satisfied.

Another way of making the same point is to note that if there are N true null hypotheses in our sample, one-tenth of them should have a marginal significance level greater than .9. If one allows for the possibility that some false nulls may nonetheless generate very high values of p, the number of true null hypotheses in the sample can be estimated as no more than 10 times the number of tested null hypotheses with marginal significance levels falling into the range [.9, 1].

The critical element of our procedure is that reviewing a large number of hypothesis tests allows us to perform our analysis using a very high marginal significance level. For any individual test it is not sensible to reject or to fail to reject the null on the basis of whether the marginal significance level is below or above .9. Even if the null hypothesis were true, we would reject 90 percent of the time, and the test would provide little information.

In this paper, however, we are interested not in the truth or falsity of any one null hypothesis but in the fraction of null hypotheses that are true. The power of our test is increased by choosing a high marginal significance level. Since the size of the cutoff is known, we adjust the number of unrejected nulls for the size to estimate the number of null hypotheses that are true. By concentrating on a range of the distribution of marginal significance levels that has been generally ignored by those researchers who have tested the individual null hypotheses, we derive an estimate of the number of true nulls that is different from—and smaller than—the one that is obtained when each is tested individually at the .05 significance level.

It would not be surprising to find that most null hypotheses tested in economics are false. After all, economists typically develop models that imply that a given parameter is nonzero and pit it as an alternative against the null hypothesis that the parameter is zero. Thus null hypotheses are formulated in such a way that it is intended that they be rejected, and nearly three-quarters of economics articles in our sample do reject their central null. Therefore, in this paper we concentrate on those hypotheses that the authors concluded were *not* rejected. We determine the fraction of these *unrejected* null hypotheses that were in fact false.

Consequently, all our statistics will be conditioned on a marginal significance level that is not less than .1. The probability of finding a marginal significance level above .9—conditional on not finding one below .1—is one-ninth. Our point estimate of the number of unrejected null hypotheses that are true is therefore no greater than nine times the number of null hypotheses with reported marginal significance levels between .9 and 1.0.

III. Data

We collected our data by reading recent issues of major economics journals to find articles in which the central null hypotheses set forth by the authors had not been rejected. We limited ourselves to empirical papers that tested substantive economic hypotheses; thus we did not include tests of the "exogeneity of the instruments" or other specification tests unless they were the principal test of a substantive economic hypothesis that was the central focus of the paper. Similarly, we did not include tests of whether, for example, the elasticity of labor supply was equal to zero unless a theory posited in the paper suggested that this value was of particular interest.

For each paper, we tried to ascertain the most central hypothesis tested and, when in doubt, chose the first test presented. Thus if an author first presented ordinary least squares results and then instrumental variables results, which she or he argued were to be preferred, we used the latter; but if the instrumental variables results were included merely to show the robustness of the findings, we used the ordinary least squares results. Occasionally a test statistic was simply reported as significant or insignificant; we were unable to make use of this information.

As much as possible we tried to conform to the author's sense of what was the single most important or reliable specification. However, the choice of test was to some extent arbitrary. This raises the possibility of "coding bias": perhaps our judgments of what was the principal hypothesis test of a paper were unduly influenced by our expectations of the results of this project. A better experimental design—one common among psychologists—would have been to have the data coded by assistants who have no point of view about or stake in the outcome of the research project.

We began by examining what we take to be the four principal journals read by American economists: the *American Economic Review*, *Econometrica*, the *Journal of Political Economy*, and the *Quarterly Journal of Economics*. After examining two years' worth of *Econometrica* and uncovering only two articles in which the central null hypothesis failed to be rejected, we substituted the *Review of Economics and Statistics* for *Econometrica*. Our data are therefore taken from the *AER* (1984–88), *Econometrica* (1986–87), the *JPE* (1984–87), the *QJE* (1985–88), and *REStat* (1986–88). The final sample consisted of 94 articles from *REStat*, 81 from the *JPE*, 73 from the *AER*, 16 from the *QJE*, and 12 from *Econometrica;* 78 of the total of 276 central hypothesis tests failed to reject the null at the .1 level.

A critical issue concerns how to treat one-sided tests. When the technique does not constrain the coefficients, the distribution of the

test statistic is the same under the null regardless of the alternative chosen. However, when the alternative suggests a particular sign for the coefficient, the econometrician is interested only in the upper or lower tail of the distribution. The question then arises whether, for coefficients that are positive under the alternative, we should calculate the probability of obtaining a coefficient that is greater than the estimated coefficient (the one-sided test) or the probability of getting a coefficient that is greater than the estimated coefficient in absolute value (the two-sided test).

We can see a case for the former procedure. However, we have chosen the latter. It is possible to argue that the two-sided significance level is preferable because we are interested in rejections of the null even in directions that economists do not presently perceive as sensible. But our primary reasons for using the two-sided significance level are pragmatic. First, relatively few tests are explicitly one-sided, but it is plausible that a substantial fraction were implicitly one-sided. Consequently, we would have to either rely on personal judgments regarding the "true" alternative, and run the risk of biasing our data, or rely on what may be an arbitrary division based on whether the test was explicitly one-sided or two-sided.

Second, under the null hypothesis the one-sided and two-sided test statistics should have the same distribution. Under the alternative, p-values should be higher in the sense of stochastic dominance when the two-sided significance level is used. Thus using the two-sided significance level will cause us to estimate that a higher fraction of null hypotheses are true and is thus a conservative approach. Indeed, in only two cases in which a test was explicitly one-sided would its p-value have been higher if we had used the one-sided cumulative distribution.

We have subsequently recognized yet a third reason to use the two-sided significance levels.[2] Although we could uncover no direct evidence for this phenomenon, it is plausible that authors would be less inclined to report actual t-values when the coefficient has the "wrong" sign. In the extreme, such a selection bias would cause all one-sided p-values to be less than .5. By using the two-sided significance levels, we avoid this potential selection problem.

While it was not our objective in collecting these data to analyze the general treatment of hypothesis testing in the economics profession, we did discover some regularities that we think are worth reporting in their own right. First, in the vast majority of cases, test statistics significant at the .1 but not the .05 level are treated as significant rejections of null hypotheses, often, but not always, justified

[2] We are grateful to our referee for forcing us to concentrate on this issue.

by the similarity of results across specifications or by the finding of a "significant" coefficient in a subsequent properly tuned specification. While this practice does not conform to the teachings of classical statisticians, it may nevertheless be sensible. Since in practice .1 appears to be the critical value for rejecting or failing to reject nulls, we treated "unrejected nulls" with marginal significance levels below .1 as rejections.

Perhaps the most striking serendipitous finding to us was the relative scarcity of formal hypothesis testing in the major journals. In the absence of *REStat* and the papers and proceedings issue of the *AER*, papers organized around formal tests of central null hypotheses would be scarce.

IV. Results

Table 1 presents the distribution of the probabilities associated with the test statistics in the papers we analyzed, along with the implied upper bounds on π, the fraction of null hypotheses that are true. The choice of the p-value p^* at which to estimate the bound is somewhat arbitrary. Higher values of p^* involve a trade-off. They are less biased if it is in fact the case that $[1 - G(p)]/(1 - p)$ is declining in p. But our estimate of the number of true nulls will be less precise.

If we choose p^* very close to one, the probability of finding even one p-value above p^* can be made arbitrarily small for any finite sample, and so our estimate of π becomes very imprecise. If we

TABLE 1

DISTRIBUTION OF REPORTED MARGINAL SIGNIFICANCE LEVELS

Marginal Significance Levels	Number of Hypothesis Tests	Estimated Upper Bound on True Nulls/ Unrejected Nulls*
1.0–.9	0	0%
.9–.8	4	23
.8–.7	7	42
.7–.6	7	52
.6–.5	6	54
.5–.4	11	66
.4–.3	11	75
.3–.2	14	86
.2–.1	18	100

* Estimated by the ratio of the number of hypotheses with marginal significance levels in this category or higher to the number of hypothesis tests that should fall in this category or higher, if all null hypotheses or all unrejected null hypotheses were true.

choose p^* further from one, our estimate becomes more precise, but it also contains a larger upward bias.

We focus attention on the tighter bounds obtained for values of p^* fixed near one at .9 and .8. The conclusions are striking. In our sample, there are no values of $f(a)$ greater than .9. Among unrejected hypotheses, one-ninth of $f(a)$ values should fall into the range .9–1.0 when the null hypothesis H_0 is true. Our point estimate of the number of unrejected nulls that are in fact true is nine times the number of marginal significance levels that fall between .9 and 1.0. *The implied estimate of* π *is therefore zero:* no null hypotheses are true.

A less extreme estimate comes from examining the fraction of unrejected null hypotheses with $f(a) > .8$. Two-ninths of unrejected nulls should fall into this category when the null hypothesis is true; we actually find that only four out of the 78 unrejected nulls (and the 276 nulls tested) do so. This produces a point estimate that 23 percent of *unrejected* null hypotheses are true.

An alternative way of approaching the issue is to assume that if each null hypothesis is true, the events $W_i = \{a_i | f(a) > .9\}$ for each hypothesis test i are independently distributed (however, there is overlap in the data used in different articles, so the independence axiom is surely false). Under the null, $P(W_i) = P(f(a_i) > .9) = .1$; under the alternative, $P(W_i) = P(f(a_i) > .9) \leq .1$. We can therefore construct a test of the super hypothesis that the unobserved fraction of all null hypotheses that are true is π or greater, for any fixed π. If more than 25 of the 78 unrejected null hypotheses in the articles in our sample are true, the odds of finding no W_i events—no cases in which $f(a) > .9$—given that $f(a) > .1$ are less than .05.

Therefore, at conventional levels of significance we can reject the hypothesis that more than 25/78, or a little less than one-third, of the *unrejected* null hypotheses in our sample are true. (Our review of hypothesis testing suggests that .1 is a more conventional level, at least for the central null hypothesis under study in an article; at this significance level we can reject the hypothesis that more than one-quarter of unrejected null hypotheses are true.)

Our failure, for $\pi < \frac{1}{3}$, to reject the null hypothesis that a fraction π of null hypotheses are true is itself an economic hypothesis. If this article is published in an economics journal, the logic of our argument would imply that this null hypothesis is also false. Without a full-blown Bayesian analysis, we cannot make precise statements about our posterior distribution over the truth or falsity of null hypotheses. It is nevertheless worth pointing out that our test does have substantial power: if more than five of the 78 unrejected null hypotheses were true, we would have less than a 50-50 chance of finding none with a marginal significance level above .9.

One item of concern is that economists do not always report significance levels, but instead simply report a test statistic as significant or insignificant. If insignificant test statistics were systematically not reported when they were extremely low, our procedure would be biased toward finding that most unrejected nulls were false. However, failure to report insignificant test statistics is in fact quite rare in the published literature. When it does occur, the typical statement is that some set of coefficients, or their sum or difference, is not significant. In these cases, it is possible to examine the t-statistics on the individual coefficients to determine whether it is likely that the test statistic for the joint hypothesis would be very low. We find no evidence of this sort of bias.[3]

A related problem is that some authors may not perform hypothesis tests, but rely on simple plots of the data as sufficiently convincing. Again, if researchers do not perform significance tests when the null is clearly not rejected, our results may reflect bias in reporting rather than the falsity of most unrejected nulls.

"Proof by plotting the data," however, is relatively rare, perhaps because "ocular regressions" are widely thought to be subject to obvious pitfalls. Two literatures that do generate "hypothesis tests" and that do not often rely on statistical methods are economic history and experimental economics. In these literatures, the hypotheses often cannot be given a statistical interpretation. "Did the winning bid in an auction converge to the theoretically anticipated outcome?" is the question, and the answer is almost invariably "no." In general, experimentalists draw judgments about whether the results are "close" to those predicted by theory given that perhaps some of the participants did not have the objective function the researcher intended them to have.

A similar problem arises in economic history. Even though one of the authors of this paper is an economic historian, we nevertheless found it difficult to determine the null hypothesis for history papers that did not present formal hypothesis tests. In part, this reflects that

[3] In the cases we recorded, we found one study in which a test statistic described as "insignificant" was a joint test of eight variables that each had a mean t-statistic of 1.44. If the coefficients were independent, then the joint p-value would be between .1 and .2. In a second case, the t-statistics on two highly correlated variables were .89 and 1.5, and their difference was characterized as "insignificant." In a third case, the author looked at the sum of the coefficients on the current and lagged exchange rate and characterized it as "insignificant." If the coefficients were independent, their sum would have a t-statistic of .27 and thus a marginal significance level less than .9. Since the coefficients on as serially correlated a variable as the exchange rate and its lag are likely to be negatively correlated, the true t-statistic is likely to be higher, perhaps significantly. Thus our review of the limited evidence at our disposal does not suggest that we have missed a number of tests because test statistics with p-values in the range [.9, 1] were not reported.

historians fall naturally into a rhetoric in which effects are judged by their substantive and economic rather than statistical significance. Thus Romer (1986) does not formally test whether the volatility of unemployment is smaller after World War II than before the Depression. Instead, she notes that the magnitude of the estimated change is not large when measured by the yardstick of the extravagant rhetorical claims for the stabilization of the postwar economy. It is not clear whether her paper would be formalized by taking as the null hypothesis that there had been no change in volatility or that the variance had fallen by more than half.

In sum, while it is possible that our failure to find many reported significance tests in which the null hypothesis fails to be rejected with a very high p-value reflects reporting bias, we do not find evidence of this in the literature. Of course, it is very possible that the bias occurs at an even earlier stage: when the data strongly support the null, the paper is less likely to be written and, if written, is unlikely to be published.

This explanation of the results is very close to the one we prefer. We therefore turn to issues of interpretation.

V. Interpretation

A rational Bayesian would use our result to draw what seem to be paradoxical inferences. On reading in a leading economics journal an article in which the central null hypothesis H_0 was not rejected, she or he would note that the sample data themselves did not appear to speak strongly against the null hypothesis. But she or he would also note that the experiment itself was drawn from a larger population—that of the subject matter of published economics articles—in which the null hypothesis is almost never true.

This prior population information—that almost all null hypotheses are false—would dominate the posterior evaluation. If in a state of relative ignorance before reading the article, after finishing the reader would be highly confident that the null hypothesis under discussion was false, even though the author of the article has failed to reject and had provisionally concluded that the null hypothesis is true. Thus there is a sharp difference between the inferences drawn from the article based only on the evidence internal to the article itself that is presented and the inferences drawn from the article taking into account both what the article explicitly says and what the existence of the article itself reveals.

One possible response to this paradox is to say that this was something economists knew all along: all null hypotheses are false, because all null hypotheses are simple shorthand descriptions of a complex

world. The key question instead is whether a null hypothesis is "good enough" for empirical work: whether the deviations between the null and the real world are sufficiently small to make conclusions reached conditional on the null reliable guides to the world or are economically significant.

There is a good deal to this argument. It is essentially an argument against hypothesis tests and for confidence intervals: economists should report not whether or not they can reject the null but whether or not their confidence interval excludes (*a*) economically insignificant values or (*b*) economically significant values. With this we agree, and we think that Leamer's (1978) and McCloskey's (1985) arguments for reorienting the rhetoric of economics toward focusing on confidence intervals have the truth on their side. Reports of empirical work should present the map the data generate from priors to posteriors and so should focus on confidence intervals and on the sensitivity of the results to small changes in specification (as in Leamer [1983] and Leamer and Leonard [1983]), even if they do not present their results within a full-blown Bayesian framework (see Zellner 1971).

It should, however, be noted that for the most part economists do not act as though they know that their hypotheses are false and are merely seeking to establish their quality as approximations. The practice of econometrics suggests that economists take their hypotheses seriously. As one example, recall that the "unit root" literature has seen a great deal of effort devoted to determining the asymptotic distribution of test statistics under the null and testing the null hypothesis that the coefficients in a univariate autoregressive model of U.S. gross national product sum to exactly one. Such a focus on the *exact* implications of what is formulated as a lower-dimensional subspace of possible parameter values for test statistics is difficult to understand if the null is viewed as only an approximation.[4]

In any event, the fact that all hypotheses are mere approximations does not completely account for our results. Economics articles are sprinkled with very low *t*-statistics—marginal significance levels very close to one—on nuisance coefficients. Very low *t*-statistics appear when the null hypothesis tested is a subsidiary one from the standpoint of the main thrust of the paper. Very low *t*-statistics appear to be systematically absent—and therefore null hypotheses are over-

[4] In fact, Christiano and Eichenbaum (1989) argue that the entire literature is badly posed because it has focused on whether or not processes contain a "unit root"; they suggest that this issue is seen as unimportant once one recognizes that the "implications of a broad class of dynamic models are reasonably robust to whether the forcing variables . . . are modeled as trend or difference stationary" (p. 23). They argue, we believe correctly, that a great deal of confusion was created by the 0-1 stationary unit root formulation of the issue.

whelmingly false—*only* when the universe of null hypotheses considered is the central themes of published economics articles.[5]

This suggests, to us, a publication bias explanation of our finding. What makes a journal editor choose to publish an article that fails to reject its central null hypothesis, which produces a value of $f(a) > .1$ for its central hypothesis test? The paper must excite the editor's interest along some dimension, and it seems to us that the most likely dimension is that the paper is in apparent contradiction to earlier work on the same topic: either others working along the same line have in the past rejected the same null, or theory or conventional wisdom suggests a significant relation.

When will there have been earlier papers along the same lines that rejected the null or strong theoretical arguments that the null is false? When the null hypothesis is in fact false. Authors therefore face a catch-22: papers that fail to reject their central null hypothesis will be published only when editors think that they are especially interesting, but editors will think that they are especially interesting only when the null hypothesis that they test really is false. Our paper can be interpreted as arguing that this social screening device is in fact quite powerful, so powerful that at most a very small proportion of failures to reject a null hypothesis can be taken at face value.

As a referee has suggested to us, a large number of papers testing for a unit root in U.S. real GNP are thought by editors to be interesting precisely because it is difficult to pin down the correct answer. It is precisely the substantial evidence the other way that makes papers that fail to reject the null of a unit root (or the null of stationarity) publishable. When a null hypothesis is well established, there is no longer space in major economics journals for papers that provide yet another failure to reject it.

Yet another alternative explanation of our results is that we have ignored a well-known fact: applied econometricians do not follow classical procedures; therefore, t-statistics are misleading and reported marginal significance levels incorrect. Most of us suspect that most empirical researchers engage consciously or unconsciously in data mining. Researchers share a small number of common data sets; they are therefore aware of regularities in the data even if they do not actively search for the "best" specification. There seems to be no practical way of establishing correct standard errors when researchers have prior knowledge of the data, or when they report only their

[5] By our count, the December 1989 *AER* contains 220 insignificant coefficients on auxiliary variables, of which 24—a little less than one-ninth—had p-values above .9. This presence of very insignificant nuisance coefficients suggests that the absence of high p-values for the central tests of empirical articles is an interesting anomaly.

favorite results; the distribution of the 10 highest t-statistics is not well known.[6]

One possible reaction is to adjust standard errors by some multiplicative factor that "compensates" for this abuse of classical procedures. Along these lines, we can use our data to ask the question, By what factor would we have to divide reported t-statistics so that one-ninth of unrejected nulls would exhibit a marginal significance level of .9 or more? The answer is about 5.5. The t-statistic of two rule of thumb would then suggest that only unadjusted t-statistics of 11 or more should be taken seriously, in which case hypothesis testing—especially in macroeconomics—would become largely uninformative. Empirical work would play only a very minor role in determining the theories that economists believe. Some claim that at present empirical work does play a very minor role in determining the theories that economists believe (see McCloskey 1985).

While we have sympathy with this reaction—and neither of us takes reported t-statistics at face value—we do not think that this is ultimately the proper road to take. While we readily believe that researchers data-mine to produce t-statistics above 1.64 or below 1.96, we see little reason to expect this bias to permeate results well outside of this range. Since neither of us sees his comparative advantage as lying in high theory, our skepticism is perhaps enhanced by the nihilistic implications regarding the role of empirical work should we set the required level of significance at an unadjusted t-statistic of 11.

VI. Conclusions

At the simplest level our findings reinforce previous calls for economists to concentrate on the magnitudes of coefficients and to report confidence levels and not significance tests. If all or almost all null hypotheses are false, there is little point in concentrating on whether or not an estimate is distinguishable from its predicted value under the null. Instead, we wish to cast light on what models are good approximations, which requires that we know ranges of parameter values that are excluded by empirical estimates.

It appears to us that a number of researchers have implicitly taken the view that explicit testing of hypotheses convinces no one, preferring to develop a "persuasive collage" of evidence. They attempt to establish a set of empirical regularities and interpret them as favorable or unfavorable to a substantive economic hypothesis. While we have some sympathy with this view, we nevertheless believe that there

[6] Especially sobering is the ease with which Hendry (1980) uses spurious variables to generate close within-sample fits and accurate beyond-sample predictions.

is a role for hypothesis testing because of the discipline it places on argument. However, hypothesis tests should concentrate on implications that are robust to minor changes in specification. Moreover, the key question should not be, Can I reject zero? Instead it should be, Can I reject all small (or all large) values for this parameter?

Our findings also pose a very peculiar epistemological problem for those interrelated literatures that have relied heavily on the failure to reject point nulls: tests of efficient markets, of the effects of anticipated variables, and of unit roots. These three literatures account for about one-third of the unrejected null hypotheses in our sample. A rational Bayesian, however, reading each paper that fails to find effects of anticipated money concludes that previous work has given the profession strong priors that anticipated money has effects and is more convinced that anticipated money does have effects, and reading each paper that fails to find profitable trading rules, is more convinced that such profitable trading rules exist. How can one do convincing empirical work in support of these null hypotheses if each published paper that fails to reject the central nulls only provides evidence to rational readers that they are false?

References

Berlin, Jesse A.; Begg, Colin B.; and Louis, Thomas A. "An Assessment of Publication Bias Using a Sample of Published Clinical Trials." *J. American Statis. Assoc.* 84 (June 1989): 381–92.

Christiano, Lawrence J., and Eichenbaum, Martin S. "Unit Roots in Real GNP: Do We Know, and Do We Care?" Working Paper no. 3130. Cambridge, Mass.: NBER, October 1989.

Hendry, David F. "Econometrics—Alchemy or Science?" *Economica* 47 (November 1980): 387–406.

Leamer, Edward E. *Specification Searches: Ad Hoc Inference with Nonexperimental Data.* New York: Wiley, 1978.

———. "Let's Take the Con out of Econometrics." *A.E.R.* 73 (March 1983): 31–43.

Leamer, Edward E., and Leonard, Herman B. "Reporting the Fragility of Regression Estimates." *Rev. Econ. and Statis.* 65 (May 1983): 306–17.

McCloskey, Donald N. *The Rhetoric of Economics.* Madison: Univ. Wisconsin Press, 1985.

Popper, Karl. *The Logic of Scientific Discovery.* New York: Harper and Row, 1959.

Romer, Christina D. "Is the Stabilization of the Postwar Economy a Figment of the Data?" *A.E.R.* 76 (June 1986): 314–34.

Zellner, Arnold. *An Introduction to Bayesian Inference in Econometrics.* New York: Wiley, 1971.