THE ACCOUNTING REVIEW
Vol. LXII, No. 1
January 1987

EDUCATION RESEARCH

Frank Selto, Editor

Inference from Empirical Research

David Burgstahler

ABSTRACT: Many researchers describe themselves as Bayesians in that they revise their prior beliefs based on observed empirical evidence. However, most studies are designed and reported as classical hypothesis tests, and research design issues are typically considered as determinants of abstract properties of statistical tests. Thus, although the primary function of empirical research is to influence beliefs, research design issues are seldom considered in their fundamental role as determinants of beliefs. In this paper, a Bayesian perspective is used to analyze the role of basic properties of hypothesis tests in the revision of beliefs. Two main points are emphasized. First, hypothesis tests with low power are not only undesirable *ex ante* (because of the low probability of observing significant results) but also *ex post* (because little probability revision should be induced even when significant results are observed). Second, irrespective of the usual issues of statistical and methodological validity, the effective level of tests in published research is likely to exceed the stated level, thus reducing the amount of probability revision justified by reported results. In combination, these conclusions are especially troublesome. If tests reported in the accounting literature are characterized by both low power and high effective levels, the results of published tests properly have little or no impact on the beliefs of a Bayesian. The Bayesian framework is useful in understanding and analyzing the tradeoffs which are an inherent part of empirical research. The analysis here identifies a Bayesian motivation for the common recommendations that researchers should attempt to maximize power in the design and execution of empirical tests and attempt to maintain the effective level of tests at their stated levels. Further, the analysis demonstrates the importance of explicit descriptions of research choices to allow (Bayesian) readers to properly revise their beliefs in response to reported empirical evidence. Finally, the model illustrates the role of prior beliefs and the characteristics of empirical tests in research and publication decisions.

THE merits of formal Bayesian methods have been discussed in numerous articles and books.[1] However, this discussion has had relatively little effect on the way research is conducted and reported in accounting and related fields.[2] Most studies are designed and reported in terms of classical hypothesis tests, and applications of formal Bayesian methods are the exception rather than the rule. Consequently, most empirical issues are analyzed only in terms

I would like to thank participants in the 1983 UBC-Oregon-Washington Research Conference and especially Robert Bowen, Lane Daley, Jim Jiambalvo, Bill Kinney, Eric Noreen, Jamie Pratt, Graeme Rankine, Ed Rice, Hein Schreuder, and three anonymous reviewers for helpful comments on earlier drafts of this paper.

David Burgstahler is Assistant Professor of Accounting, University of Washington.

Manuscript received July 1985.
Revisions received April 1986 and June 1986.
Accepted July 1986.

of the effect on properties of classical hypothesis tests. The more fundamental issue of the effect on belief revision is not directly addressed.

Although formal Bayesian methods are seldom used, many accounting researchers describe themselves as Bayesians who revise their prior beliefs based on observed empirical evidence. The purpose of this paper is to explore the implications of a Bayesian approach to revising beliefs based on the reported results of classical hypothesis tests. This framework provides useful insights into the process by which empirical evidence is used to revise beliefs. For example, the Bayesian perspective demonstrates the distinction between statistical significance and statistical persuasiveness by showing that there are situations where statistically significant results should not lead to revision of the reader's prior beliefs. Further, the characteristics of the research and publication process suggest that these situations may not be unusual.

The remainder of the paper is organized as follows: After briefly reviewing the fundamentals of classical hypothesis testing in the following section, a Bayesian framework for integrating empirical evidence with prior beliefs to form posterior beliefs is described in the next section.[3] In this framework, three factors jointly determine posterior beliefs. The first two, the power and significance level of the empirical test, are discussed in the succeeding two sections. The third factor, prior beliefs, is discussed next along with other factors which influence research and publication decisions. Finally, a summary and conclusions are presented.

Hypothesis Testing

Development of a hypothesis test comprises three interrelated steps: (1) choose a test statistic, (2) derive the distribution of the test statistic under the null hypothesis, and (3) define a rejection region such that the probability of observing a test statistic in the rejection region if the null hypothesis holds is some (prespecified) significance level. If the observed value of the test statistic falls in the rejection region, the null hypothesis is rejected; otherwise, the null hypothesis is not rejected. Since the probability of the statistic falling in the rejection region if the null hypothesis holds is relatively small (i.e., equal to the level of significance), observation of a significant statistic is interpreted as evidence against the null hypothesis.

The preceding description omits some important aspects of the problem of designing a satisfactory hypothesis test, but the scope of published descriptions of hypothesis tests (and, by implication, the scope of issues considered in constructing tests) is often similarly limited. In developing and describing a hypothesis test, attention is typically focused on the significance level of the test. However, if the level of significance were the only concern, a random number generator could be used to construct an essentially costless test of any hypothesis.

Another critical concern is the power of the test.[4] The alternative hypothesis implies the distribution for the test statistic which determines the power of the

---

[1] See, for example, Savage [1954], Raiffa and Schlaifer [1961], Edwards, Lindman, and Savage [1963], or Zellner [1971].

[2] Efron [1986] suggests several factors which may explain why most scientific data analysis is carried out in a non-Bayesian framework.

[3] Similar models are found in Zellner [1971, Chapter 10] and Judge et al. [1985].

[4] In the larger context of a decision problem, the loss function must also be viewed as an integral part of hypothesis testing (see Savage [1954, Chapter 16]).

test, i.e., the probability that the test statistic will fall in the rejection region (will be significant) when the alternative hypothesis holds.[5]

In evaluating results of completed empirical research, there tends to be an emphasis on either the null or the alternative hypothesis distribution, but not both. When a significant test statistic is observed, factors which might have caused the effective level of the test to exceed the stated level are explored. On the other hand, when the observed test statistic is not significant, factors which might have caused the power of the test to be low are examined. However, as emphasized in the following sections, both the level and power of a test are important *regardless of the outcome of the test*, i.e., regardless of whether the observed statistic is significant.

A Bayesian Framework for Integrating Empirical Evidence

Let H_0 and H_A, respectively, denote the null and alternative hypotheses. Let the observation of a significant test statistic be denoted by S and observation of a nonsignificant test statistic by NS. Finally, let the probability of observing S when the null hypothesis holds be α and let the probability of observing NS when the alternative holds be β. Then the following familiar table of probabilities of observing S or NS conditional on H_0 or H_A applies:

Conditional Probabilities of Outcomes

		State	
		H_0	H_A
Value of Test Statistic	NS	$P[NS\|H_0]=1-\alpha$	$P[NS\|H_A]=\beta$
	S	$P[S\|H_0]=\alpha$	$P[S\|H_A]=1-\beta$

A Bayesian will combine prior beliefs about H_0 with observed empirical evidence to form a posterior belief. For example, the posterior belief in H_0 given that empirical evidence S is observed is:

$$P[H_0|S]=\frac{P[S|H_0]P[H_0]}{P[S|H_0]P[H_0]+P[S|H_A]P[H_A]}=\frac{\alpha P[H_0]}{\alpha P[H_0]+(1-\beta)P[H_A]} \quad (1)$$

where $P[H_0]$ and $P[H_A]$ denote the prior beliefs in H_0 and H_A, respectively. Similarly,

$$P[H_A|S]=\frac{(1-\beta)P[H_A]}{\alpha P[H_0]+(1-\beta)P[H_A]}=1-P[H_0|S] \quad (2)$$

In the following sections, the posterior belief in the null hypothesis given observation of a significant statistic, $P[H_0|S]$, is discussed as a function of the power of the test $(1-\beta)$, the level of the test (α), and the prior belief in the null $(P[H_0])$.[6]

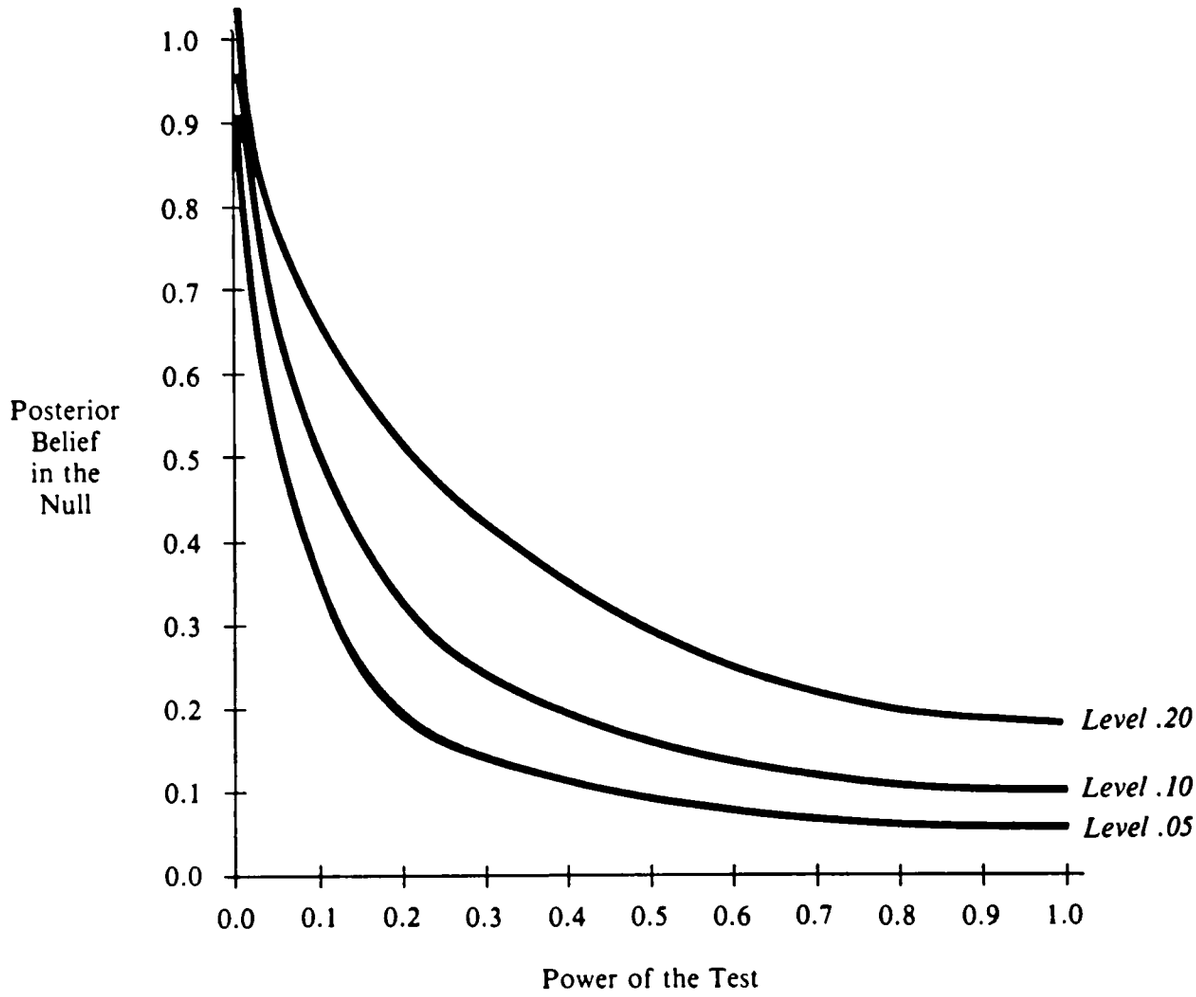
Power as a Determinant of Posteriors

Although the issue of power is often discussed when a nonsignificant test sta-

[5] Composite alternative hypotheses would imply a set of alternative hypothesis distributions, giving rise to a power function defined over the set of alternative hypotheses. Since composite alternative hypotheses complicate the exposition without changing the qualitative conclusions, only simple alternatives are considered here. For an application of the model in Section 3 to composite alternatives, see Zellner [1971].

[6] Since $P[H_0|S]=1-P[H_A|S]$, discussion of the posterior belief in the alternative hypothesis would be logically equivalent.

FIGURE 1
POSTERIOR BELIEF IN THE NULL AFTER OBSERVING A SIGNIFICANT TEST STATISTIC
(FOR PRIOR BELIEF OF .5)



tistic is observed, the importance of power when a significant statistic is observed has received little attention.[7] The power of a test which has resulted in a significant test statistic is commonly viewed as a moot issue; there seems little reason to be concerned with the power of a test which has been "powerful enough" to yield a significant test statistic. However, equations (1) and (2) show that power $(1-\beta)$ remains a critical component of the belief revision process even when a significant test statistic is observed. A Bayesian cannot know how (or even whether) to revise beliefs in response to a significant test statistic without knowledge of the power of the test.

This point can be further illustrated by examining posterior beliefs given observation of a significant test statistic as a function of power. Figure 1 plots the relationship between posterior belief in the null hypothesis and power of the test for three different significance levels,

[7] For example, Cook and Campbell [1979, p. 40] emphasize the importance of analysis of power when no effect is observed but do not mention the importance of power when a significant effect is observed. See also Simonds and Collins [1978, pp. 649-650], Foster [1980, p. 41], Foster [1981, pp. 220-222], Ball and Foster [1982, p. 186], Beaver [1982, pp. 327-328], and Kinney [1986, pp. 345-348].

.05, .10, and .20, with prior beliefs $P[H_0] = P[H_A] = .5$.[8] For more powerful tests, there is a lower degree of belief in the null (and greater belief in the alternative) as a result of observation of a significant statistic.[9] If the power of a test is as low as the significance level of the test, then the posterior will be unchanged from the prior *regardless of the outcome of the test.*[10] This is illustrated in Figure 1 where posterior beliefs are seen to be equal to the prior belief (.5) at the value where power is equal to the level of the test. Intuitively, when a significant test statistic is equally likely under the null and alternative hypotheses (i.e., when the level and power of the test are the same), results of the test do not change beliefs about the null hypothesis.[11]

In empirical accounting research, decisions made by researchers often result in bias toward non-rejection of the null hypothesis, i.e., they reduce the power of the test.[12] The *ex ante* risk that the reduction in power will result in insignificant (and probably unpublishable) results is well-known and widely acknowledged. However, it is not widely recognized that this is more than an *ex ante* problem. Even if the results from a test with low power are significant, the results have little value; even significant results from a low-power test do not cause much change in a Bayesian's beliefs about the null hypothesis.

Researchers sometimes expect a single set of data to serve two purposes. Results are calculated to decide whether a test is sufficiently powerful as well as to draw a conclusion about the hypothesis being tested. In fact, researchers may even mistakenly assert that a significant result from a low-power test is more convincing evidence against the null than a significant result from a high-power test because a more extreme test statistic is required to attain significance for a low-power test.[13] However, not all tests which yield significant results are powerful tests and it is inappropriate to judge the power of a test based solely on the significance of the observed results. Moreover, from equations (1) and (2) it is clear that the power of the test is a required input to Bayesian belief revision in response to empirical results.

In summary, the power of a test is critical in drawing inferences from empirical results. The statement that significant results are evidence against the null hypothesis is, by itself, incomplete; significant results are only strong evidence against the null for powerful tests.

Effective Level of Published Empirical Tests

The role of the effective level of pub-

[8] The specific significance levels and prior beliefs plotted in Figure 1 were chosen to provide quantitative examples, but similar qualitative conclusions would apply for any (non-dogmatic) priors and any (non-zero) significance levels.

[9] Note, however, that even for a test with power of 1.0, a significant statistic does not result in a posterior belief in the alternative of 1.0 unless the level of the test is 0.0. An ideal test with power 1.0 and level 0.0 would enable the researcher to reason by logical implication rather than by statistical inference (see Bakan [1966]).

[10] When the power of the test is less than the effective level of the test, observation of a "significant" test statistic actually results in an *increased* belief in the null hypothesis. In this somewhat perverse situation, referred to as a biased hypothesis test [Mood, Graybill, and Boes, 1974], a significant test statistic is actually more likely under the null than under the alternative and thus increases belief in the null.

[11] Thus, a "significant" test statistic from the random number generator mentioned earlier would not change beliefs because a significant test statistic would be equally likely under the null and alternative hypotheses.

[12] For some specific examples, see Beaver, Clarke, and Wright [1979, pp. 326–327], Beaver, Lambert, and Morse [1980, pp. 11–12], and Bowen, Noreen, and Lacey [1981, p. 178].

[13] For example, Zmijewski and Hagerman [1981, p. 138] construct a test so as "to intentionally bias the estimator in favor of the null hypothesis to provide for a stronger test." Similarly, Bakan [1966] concludes that a significant result from a test with a small sample size is more convincing than a significant result from a test with a larger sample size.

lished tests in belief revision is illustrated by the three curves in Figure 1 which show posterior beliefs for tests of three different significance levels. Observation of a test statistic significant at a lower level, α, results in a lower posterior belief in the null hypothesis, $P[H_0|S]$. Thus, the effective level of a published result is a critical component of the belief revision process for a Bayesian.

The publication process serves as a filter which affects the probability of observing a published significant test statistic when the null hypothesis holds.[14] Some of the incentives which determine the characteristics of this filter are discussed in the following section. In this section, behavioral factors which may cause the effective level of the tests reported in the literature to exceed their nominal level are discussed.[15] These behaviors create biases which are described below as researcher-induced or editor-induced, though the categories are not always distinct. Researchers may be aware of many of these biases individually but their joint and cumulative effects are not often considered, nor is their impact on a Bayesian integration of empirical evidence with prior beliefs.

Researcher-induced biases are the result of choices made in the course of a research project. Greenwald [1975, p. 3] summarizes a number of researcher behaviors which lead to a high probability of observing significant results in published research even when the null hypothesis holds (i.e., a high effective level for published tests), including:

1. submitting results for publication more often when the null hypothesis has been rejected than when it has not been rejected;
2. continuing research on a problem when results have been close to rejection of the null hypothesis ('near significant'), while abandoning the problem if rejection of the null hypothesis is not close;
3. elevating ancillary hypothesis tests or fortuitous findings to prominence in reports of studies for which the major dependent variables did not provide a clear rejection of the null hypothesis;
4. revising otherwise adequate operationalizations of variables when unable to obtain rejection of the null hypothesis and continuing to revise until the null hypothesis is (at last!) rejected or until the problem is abandoned without publication;
5. failing to report initial data collections (renamed as 'pilot data' or 'false starts') in a series of studies that eventually leads to a prediction-confirming rejection of the null hypothesis; and
6. failing to detect data analysis errors when an analysis has rejected the null hypothesis by miscomputation, while vigilantly checking and rechecking computations if the null hypothesis has not been rejected.

The effective significance level of published results may also be increased by a practice sometimes described as "refining a theory" in the course of a research project. For example, a project might begin with a theory which suggests the direction of a relationship between two or more variables but is not sufficiently refined to specify the precise functional form of the relationship between the dependent variable and the independent variables. The theory may not be sufficiently well-developed to specify which variables should be included, whether the relationship should be linear or

[14] Further discussion of this point is found in Sterling [1959], Bakan [1966], Lykken [1968], Walster and Cleary [1970], Greenwald [1975], and McCloskey [1983].

[15] The discussion in this section focuses on some behavioral factors which may lead to invalid nominal significance levels. A more general discussion of threats to validity is found in Cook and Campbell [1979].

quadratic, or whether the effects of independent variables are additive or multiplicative. As the research progresses, the theory is sometimes "refined" by trying several functional forms and selecting the one which is most consistent with the data. For example, Bowen [1981] explicitly uses significance and consistency with theory as criteria for choosing a particular functional form for a regression equation.[16] When data-fitting is used to refine the form of a regression equation and the same data are used for hypothesis tests, the effective significance level of the test may substantially exceed the stated level.[17]

The most troublesome aspect of the behaviors discussed above is that their precise effects on the effective level of published tests are difficult to identify and even more difficult to quantify. However, it would be naïve to conclude that because the effects cannot easily be quantified, they must be unimportant. Rather, researchers must attempt to minimize these effects by meticulous adherence to carefully planned research designs. Researchers should begin with a well-developed theory which specifies the relevant variables and the functional form of the relationship. Whenever possible, major research decisions such as operational definitions of variables, data to be collected, and sample sizes should be made in advance and researchers should be reluctant to modify the research plan unless the modified plan is viewed as a separate test requiring new data. Only by conducting carefully planned tests of well-developed theories can researchers effectively control the significance level of reported tests.

The proportion of incorrect rejections in published resarch is also increased by the behavior of journal editors and reviewers. If, as casual empiricism suggests, stricter editorial standards are applied to completed research with nonsignificant results, the proportion of published incorrect rejections of the null hypothesis will be higher than implied by the nominal level of the tests.[18] In fact, if only rejections of null hypotheses were published, then, even for a correct null hypothesis, the probability of observing a published study which rejects the hypothesis would increase rapidly as the number of tests of the hypothesis increased.[19]

Most researchers conscientiously attempt to avoid the behaviors described above, though few can claim to be free of their effects. Also, although there is probably some editorial bias against publishing insignificant results, editors generally do not follow a policy of simply publishing all significant results and rejecting all nonsignificant results. Well-designed and executed research is pub-

[16] While other researchers in accounting undoubtedly consider significance and consistency in selecting a model to be used for significance tests, Bowen is especially conscientious in explicitly reporting his criteria. When specifications are tried and discarded without being reported, the reader has no way to assess the impact of alternative specifications on the effective level of the reported tests. For other specific examples, see McCloskey's [1985] description of a sample of ten recent papers from economics using regression analysis, of which only two do not admit experimenting with alternative specifications.

[17] Freedman [1983] demonstrates how seriously the effective significance level can be distorted by selecting variables for inclusion in a regression equation based on significance of the individual coefficients.

[18] In order to eliminate the bias in acceptance criteria, Walster and Cleary [1970] suggest an editorial policy in which data and results are withheld when an article is submitted for review so that publication decisions are based on the design rather than the results. Kinney [1986] suggests a similar policy for evaluating dissertation research.

[19] To make the risk of observing a significant result purely by chance more concrete, let k be the number of independent tests conducted, all at level α. Then, the probability of observing at least one rejection of the hypothesis is $1-(1-\alpha)^k$. As k goes from 1 to 5 to 10, the probability of observing at least one rejection goes from .05 to .22 to .40 for $\alpha=.05$ and from .10 to .40 to .65 for $\alpha=.10$.

lished despite statistically insignificant results; poorly-designed and executed research is rejected despite statistically significant results.

Nonetheless, the conclusion remains: If any of these behaviors is exhibited by researchers or editors, the effective proportion of published incorrect rejections of null hypotheses may systematically exceed the proportion implied by the nominal level of the tests conducted. As a consequence, the belief revision by a Bayesian in response to a published empirical result should be less than would be implied by the (incorrect) nominal level of significance. Moreover, readers attempting to integrate published results with their prior beliefs cannot determine the effects of unrecorded choices made by researchers and editors and thus are missing a critical input to the belief revision process.[20, 21]

Prior Beliefs and Incentives for Research Production and Dissemination

The role of prior beliefs in belief revision is clear from equations (1) and (2); posterior beliefs are directly related to prior beliefs and individuals with different prior beliefs will have different posterior beliefs. In this section, the discussion focuses on another aspect of prior beliefs, the role of prior beliefs in research and publication decisions. Prior beliefs influence decisions because of their role in researchers' perceptions of the probable outcome of proposed research projects and in editors' interpretations of submitted results. The influence of prior beliefs is complex; researchers' decisions reflect perceptions of editors' beliefs and editors' decisions are influenced by perceptions of researchers' beliefs.

A test of a hypothesis must satisfy two necessary conditions to be considered for publication: The hypothesis (1) must not be trivial, and (2) must not be quasi-certain. A null hypothesis not related to a substantive economic issue is trivial. A null hypothesis for which substantially all researchers hold prior beliefs "close to" zero or for which substantially all researchers hold prior beliefs "close to" one is quasi-certain. The results of a test of a quasi-certain hypothesis are not informative because they are "close to" perfectly predictable.[22] The tastes and preferences of an editor determine how these definitions are operationalized. The results of tests of hypotheses which an editor deems trivial or quasi-certain will not be published. In the remainder of this section, all hypotheses contemplated by researchers and editors are assumed to meet these two necessary conditions.

Editors may be reluctant to publish significant results when prior beliefs are strong, i.e., when the prior belief in the null hypothesis is either very low or very high. When the prior belief in the null

[20] Based on a number of assumptions about researcher and editor behaviors, Greenwald [1975] estimated that the proportion of incorrect rejections of null hypotheses appearing in the social psychology literature may be as high as .30 (for tests conducted at the .05 level). Although this estimate does not necessarily apply to research in accounting (since different behavioral assumptions would lead to different estimates of the proportion of incorrect rejections), it does give an impression of the potential magnitude of the effects of these biases.

[21] The level of published tests remains a concern even with publication of the results of more than one test of a hypothesis. Christie [1985] proposes an interesting technique for combining evidence from separate tests of a hypothesis. Since the technique requires valid assessments of effective significance levels for the individual tests, correct nominal significance levels are essential.

[22] It might seem that test results inconsistent with highly certain priors would be especially interesting. However, as illustrated in the following paragraph, unless it can be established that the test had very high power and a very low significance level, it may be more likely that the results are incorrect than that the prior beliefs are incorrect.

is low, significant results are consistent with the priors and may not change beliefs much.[23] An editor may be reluctant to expend resources to publish results that merely confirm strong prior beliefs. On the other hand, when the prior belief in the null is high, an editor may also be reluctant to publish significant results. For example, with a prior belief in the null of .95, significant results from a test with a level .10 and power of .90 would cause posteriors to be revised downward to approximately .68, a fairly substantial revision of beliefs. However, the posterior belief in the null hypothesis given a significant test statistic can also be interpreted as the probability that the significant results are "incorrect." Thus, the posteriors indicate that the odds are greater than two to one that the significant result is "incorrect!" Because editors are likely to be reluctant to publish a result that has a high probability of being incorrect, hypotheses for which priors are strong require especially high-power tests with carefully controlled significance levels.

A primary purpose of publishing the results of empirical research is to influence researcher beliefs about a hypothesis and thus influence the further development of theory. However, a secondary effect, apparent from equation (3) below, is that changes in beliefs affect the expected utility to other researchers of devoting resources to tests of a hypothesis. Thus, publication of results which are largely consistent with prior beliefs can be valuable when the results strengthen beliefs so that other researchers will not choose to expend additional resources to conduct tests of the hypothesis.[24]

A researcher will choose the research project, P, which maximizes expected utility, $E(U^P)$,[25]

$$\begin{aligned} E(U^P) &= P(H_0)E(U^P|H_0) + \\ &\quad P(H_A)E(U^P|H_A) \\ &= P(H_0)\{(1-\alpha)U[H_0,NS] + \\ &\quad \alpha U[H_0,S]\} + \\ &\quad P(H_A)\{\beta U[H_A,NS] + \\ &\quad (1-\beta)U[H_A,S]\}. \end{aligned} \tag{3}$$

The expected utility of a proposed project depends on prior beliefs about the hypothesis as well as the power and level of the proposed hypothesis test. Different researchers may have different prior beliefs and may have access to tests with different powers, or have different perceptions of the power of a given test. These differences, along with other factors, lead different researchers to choose different research projects.[26]

The expected utility of a proposed project also depends on the utility of the state/outcome combinations. Publications are major determinants of the decisions which determine most monetary and nonmonetary rewards for academic researchers, including promotion and tenure decisions, salary increases, consulting rates, and job changes. Consequently, the utility of each state/outcome combination depends primarily on the product of the reward to publication and the probability of publication. Publications that deal with more important

[23] The extreme case is the test of a hypothesis which is obviously false; significant results from a test of such a (quasi-) certain hypothesis do not change beliefs at all.

[24] A third purpose of publication is to support and disseminate methodological innovation which improves the level and power of subsequent tests, thus increasing the belief revision supported by the results of subsequent research.

[25] More generally, researchers and editors must choose a set of research projects rather than just a single research project. Choosing a portfolio of projects which maximizes expected utility is a more complex problem than the one considered here and raises additional issues beyond the scope of this discussion.

[26] The utility-maximizing P^* may be the null research project which allows a researcher to engage in alternative activities, e.g., consulting or consumption of leisure.

hypotheses and that have a larger impact on beliefs are likely to lead to greater rewards to the researcher. The probability of publication depends on researcher perceptions of editorial policies and beliefs.

If, as suggested in the previous section, researchers perceive that significant results are much more likely to be published, then researchers will choose the project which yields an optimal combination of probability of significant results and low cost. Assuming that the utility for significant results ("rejecting the null") is much higher when the null is false, researchers will prefer powerful tests of null hypotheses for which they have low prior beliefs. Thus, the utility-maximizing project may be a powerful test of a hypothesis which the researcher believes is likely to be false. However, powerful tests of hypotheses are often costly; for example, a high-power test may require a large amount of researcher time for data collection and analysis or may require that the researcher acquire additional expertise. Consequently, the utility-maximizing research project could also be a low-power, but low-cost, test of a null hypothesis which the researcher believes is likely to be false.

Under some circumstances, researchers could also have high utility for rejecting a true hypothesis. For example, researchers may have high utility for publishing significant, but incorrect, results if subsequent results which contradict the original results are not likely to be published until after some decision affected by publication (e.g., tenure) has been made. Alternatively, for very costly tests or tests for which it is difficult to assess power, there may be little chance that a replication contradicting the results will ever be conducted and published to reveal an error. Even if subsequent tests reveal that the hypothesis is true, the impact on the original researcher's reputation and wealth may be small as long as there is no evidence of deceit or incompetence. Thus, because of the possibility of observing significant results even when the null hypothesis holds, the utility-maximizing research project might be a low-cost test of a null hypothesis which the researcher believes is likely to be true. In fact, in choosing a project from the set of research projects for which the power of the test is nearly as low as the level, the primary consideration would be the cost of available tests, rather than researcher beliefs about the hypotheses.

Prior beliefs play an important role throughout the research process. They directly determine posterior beliefs as shown in equations (1) and (2). However, they also play an important role in the decisions of editors and researchers. Further, the power and level of available tests and the costs associated with each test are determinants of researcher decisions. The model presented in this section helps to clarify the relationships among these factors in research decisions.

Conclusions

Research practices which lower the power of tests have sometimes been accepted in accounting research without consideration of their implications. Similarly, the implications of research and editorial practices which increase the effective level of reported tests have received little attention. The model discussed in this paper demonstrates that tests with low power are not only undesirable *ex ante*, because of the low probability of observing significant results, but also *ex post*, because little probability revision should be induced by the results of low-power tests regardless of whether significant results are observed.

Further, the discussion suggests that, irrespective of the usual issues of statistical and methodological validity, the effective level of tests in published research is likely to exceed the stated level, thus reducing the amount of belief revision justified by reported results.

The Bayesian framework presented here is useful in analyzing and understanding the trade-offs which are an inherent part of empirical research. The model clarifies the relationship between research decisions and characteristics of tests, prior beliefs, and individual utilities. Broadly stated, the recommendations which follow from the analysis are little different from those of an introductory discussion of experimental design: Researchers should attempt to maximize power in the design and execution of empirical tests and attempt to maintain the effective level of tests at their stated levels. However, the analysis identifies the reasons for these recommendations. The results of tests with low power have little value (even if the results are statistically significant) because low-power tests should induce little belief revision regardless of their outcome. Behaviors which increase the effective level of published empirical tests are undesirable because an increase in the effective level is accompanied by a decrease in the belief revision justified by the results. If these recommendations were not followed, published empirical results would properly have little or no effect on the beliefs of accounting researchers.

REFERENCES

Bakan, D., "The Test of Significance in Psychological Research," *Psychological Bulletin* (December 1966), pp. 423–437.

Ball, R., "Discussion of Accounting for Research and Development Costs: The Impact on Research and Development Expenditures," *Journal of Accounting Research* (Supplement, 1980), pp. 27–37.

———, and G. Foster, "Corporate Financial Reporting: A Methodological Review of Empirical Research," *Journal of Accounting Research* (Supplement, 1982), pp. 161–234.

Beaver, W., "Discussion of Market-Based Empirical Research in Accounting: A Review, Interpretation, and Extension," *Journal of Accounting Research* (Supplement, 1982), pp. 323–331.

———, R. Clarke, and W. Wright, "The Association Between Unsystematic Security Returns and the Magnitude of Earnings Forecast Errors," *Journal of Accounting Research* (Autumn 1979), pp. 316–340.

———, R. Lambert, and D. Morse, "The Information Content of Security Prices," *Journal of Accounting and Economics* (March 1980), pp. 3–28.

Bowen, R., "Valuation of Earnings Components in the Electric Utility Industry," THE ACCOUNTING REVIEW (January 1981), pp. 1–22.

———, E. Noreen, and J. Lacey, "Determinants of the Corporate Decision to Capitalize Interest," *Journal of Accounting and Economics* (August 1981), pp. 151–179.

Box, G., "Sampling and Bayes Inference in Scientific Modelling and Robustness," *Journal of the Royal Statistical Society—Series A* (1980), pp. 383–430.

Christie, A., "Evaluation of the Consistency of Evidence on Contracting and Political Cost Hypotheses Across and Within Studies," Unpublished paper, School of Accounting, University of Southern California (February 1985).

Cook, T., and D. Campbell, *Quasi-Experimentation: Design and Analysis Issues for Field Settings* (Houghton Mifflin, 1979).

Edwards, W., H. Lindman, and L. Savage, "Bayesian Statistical Inference for Psychological Research," *Psychological Review* (1963), pp. 193–242.

Efron, B., "Why Isn't Everyone a Bayesian?", *The American Statistician* (February 1986), pp. 1–5.

Foster, G., "Accounting Policy Decisions and Capital Market Research," *Journal of Accounting and Economics* (March 1980), pp. 29–62.

———, "Intra-industry Information Transfers Associated with Earnings Releases," *Journal of Accounting and Economics* (1981), pp. 201-232.

Freedman, D., "A Note on Screening Regression Equations," *The American Statistician* (May 1983), pp. 152-155.

Greenwald, A., "Consequences of Prejudice Against the Null Hypothesis," *Psychological Bulletin* (January 1975), pp. 1-20.

Judge, G., W. Griffiths, R. Hill, H. Lutkepohl, and T. Lee, *The Theory and Practice of Econometrics*, 2nd Ed. (John Wiley & Sons, 1985).

Kinney, W., "Empirical Accounting Research Design for Ph.D. Students," THE ACCOUNTING REVIEW (April 1986), pp. 338-350.

Lykken, D., "Statistical Significance in Psychological Research," *Psychological Bulletin* (1968), pp. 151-159.

McCloskey, D., "The Rhetoric of Economics," *Journal of Economic Literature* (1983), pp. 481-517.

———, "The Loss Function Has Been Mislaid: The Rhetoric of Significance Tests," *American Economic Review* (1985), pp. 201-205.

Mood, A., F. Graybill, and D. Boes, *Introduction to the Theory of Statistics*, 3rd Ed. (McGraw-Hill, 1974).

Raiffa, H., and R. Schlaifer, *Applied Statistical Decision Theory* (Graduate School of Business Administration, Harvard University, 1961).

Savage, L. *The Foundations of Statistics* (John Wiley & Sons, 1954).

Simonds, R., and D. Collins, "Line of Business Reporting and Security Prices: An Analysis of an SEC Disclosure Rule: Comment," *The Bell Journal of Economics* (Autumn 1978), pp. 646-658.

Sterling, T., "Publication Decisions and Their Possible Effects on Inferences Drawn from Tests of Significance—or Vice Versa," *Journal of the American Statistical Association* (March 1959), pp. 30-34.

Walster, G., and T. Cleary, "A Proposal for a New Editorial Policy in the Social Sciences," *The American Statistician* (April 1970), pp. 16-19.

Zellner, A., *An Introduction to Bayesian Inference in Econometrics* (John Wiley & Sons, 1971).

Zmijewski, M., and R. Hagerman, "An Income Strategy Approach to the Positive Theory of Accounting Standard Setting Choice," *Journal of Accounting and Economics* (1980), pp. 129-149.