Confirmation and Replication in Empirical Econometrics: A Step Toward Improved Scholarship

William G. Tomek

The strength of agricultural economics rests on its capacity to combine theory, quantitative methods, and data to do useful analyses of problems faced by society. There is a growing awareness, however, that agricultural economists are not, in fact, doing this very well (Bonnen 1988, 1991). One component of the problem is that econometric results are often fragile: small changes in the model or data result in large changes in empirical results (Leamer) and different models of the same phenomenon can give conflicting results (Hendry and Richard). Consequently, the value of applied econometrics as a tool for analyzing problems or aiding decision makers is reduced.

Econometric models are inherently approximations, and there are no easy solutions for the problem of unstable empirical results. One aid is to build more carefully on prior research, and empirical work should aim to reduce (or explain) conflicting conclusions. Researchers should demonstrate precisely how their work improves on and adds to earlier research. Replication and confirmation, I shall argue, are often essential components in demonstrating such improvements. Attempts to confirm prior work provide a depth of understanding that is otherwise unattainable.

Replication and confirmation are difficult, however, and benefits are often perceived to be smaller than costs. It is important that both the benefits and difficulties of such research be understood. Thus, one objective of this essay is to discuss the meaning and benefits of confirmation and replication, and a second is to describe the problems, hence the costs. I conclude with suggestions for encouraging improved scholarship in econometric analysis, recognizing that changes in current research protocols will not be easy. Among other things, it will require giving more weight to the quality than to the quantity of research—a criterion often espoused, but rarely honored.

Concepts and Terminology

The social mechanisms of doing scientific work include full and open communication of research methods and results, thereby permitting appraisal and extensions of the research. For experimental results, it is in principle possible to conduct independent repetitions of the experiment; the statistical model is refitted to new data obtained from the new experiments. Empirical results will vary, if for no other reason than experimental (sampling) error. When nonexperimental data are employed, as is common in economics, it is perhaps more accurate to use the terms confirmation or duplication. Most analyses of models published by others have involved attempts to duplicate results.

It is true, however, that some economists use replication to mean duplicating published results (e.g., Dewald, Thursby, and Anderson). On the other hand, Mittelstaedt and Zorn define four kinds of replication based on a two-by-two matrix, with "original" and "new" categories for both the data and the models. In my view, the categories involving new model specifications are not replication, and I think that it is useful to distinguish between the cases of duplicating published results and fitting the original specification to new data.

Thus, in this paper, I use "confirmation" or, for diversity, "duplication" to mean attempts to fit the original model with the original data. I reserve "replication" for fitting the original specification to new data. But in economics, the

William G. Tomek is a professor of agricultural economics at Cornell University.

James T. Bonnen, Curtis Braschler, Ben C. French, George W. Ladd, Anya M. McGuirk, Bernard F. Stanton, and Fred C. White made helpful comments on earlier drafts of this paper. I also acknowledge suggestions received from seminar participants at Pennsylvania State University. I am, of course, responsible for the final content. An early version appears as Cornell Agricultural Economics Staff Paper 92–1.

Amer. J. Agr. Econ. 75 (October 1993): 6–14

new data are usually not generated by formal, independent experiments. In this sense, the use of the term replication has a different meaning than in experimental sciences.

Economic activity can be viewed, however, as an uncontrolled experiment, which is recorded in a variety of data series. Conceptually, history can be repeated with the same initial conditions and exogenous factors, but the observations on the endogenous variables will vary for each rerun of history (experiment) because of different random events. In practice, of course, repeating history is impossible, and the data are usually treated "as if" they came from a single experiment.

The attempt to find better models is complicated by two unknowns. One unknown is the true model and whether structural change has occurred with the passage of time. This issue is not addressed directly in the present paper, though, of course, one of the potential benefits of confirmation research is improved models. (A definition of a statistical generating process and some related difficulties of econometric modeling are discussed in Spanos.) It also should be noted that uncertainty about the correct model leads to searches for the preferred model using one data set, i.e., pretesting.

A second, rather subtle, unknown is the conceptual nature of data revisions, and since data revisions can have rather profound effects on empirical results, the nature of these revisions needs to be better understood. If the data are viewed as being generated by an experiment of nature (the economy), then the data revision framework depends on conceptualizing the recording of observations from this experiment.

Two frameworks have been proposed (e.g., Patterson and Heravi). One is the measurement error hypothesis. It takes the view that, while information for constructing the true series is available, it is too costly for a government agency to record all available information contemporaneously with its occurrence. Thus, analogous to adjustment cost arguments for distributed lag effects, time is required to compile full (or better) information and to publish revisions. A measurement error is created because a difference exists between the actual information available to decision makers and the information contained in the preliminary data reported by government agencies. For instance, consumers presumably have good estimates of their income, but the full information about personal income is not reflected in initial estimates by the Department of Commerce.

A second view of data revisions is analogous to the notion of weak-form efficiency in markets; namely, the hypothesis is that the current vintage of the data fully incorporates all of the information available at the time the data are compiled and that subsequent revisions incorporate truly new information which was not available earlier. In this case, the preliminary estimates are unbiased forecasts of the revised series. In principle, it is possible to test which of these views correctly characterizes a particular revised series (Patterson and Heravi).

In agricultural economics, both concepts are applicable, depending on the variable. Preliminary estimates of crop size can, at least in principle, be unbiased forecasts of the final estimate. However, the measurement error hypothesis probably is the more correct of the two in most applications. For example, changes in meat processors' practices in trimming bone and fat from beef carcasses are readily apparent to buyers at the time purchases are made, but are not instantly reflected in the factor used by the USDA to compute retail beef consumption from beef carcass weights. Thus, the preliminary estimates are likely in error relative to the values used by decision makers.

The practical difficulties created by data revisions are discussed below, and the errors-in-variable framework is implicit in the discussion. But those conducting replication analyses must decide on the appropriate concept for their work. Errors in variables may imply biased estimates of the unknown parameters, and revisions presumably reduce bias. For the unbiased forecast view of data revisions, the different vintages can be treated as different samples, and the results perhaps are more like experimental replications. But whatever the nature of data revisions, experimental controls are not involved, and revisions can change the collinearity among variables, thereby changing the precision of the estimates.

To summarize, a distinction can be made between confirmation (duplication) and replication. Most "replication" research in economics has involved attempts to confirm published results. Econometric models can be refitted either to revised data for the same sample period or to data for a new sample period. This is a type of replication, but unlike experimental sciences, changes in coefficients and their variances reflect changed collinearity among variables and perhaps reduced errors in variables. Coefficients also may change because of changes in the magnitude of the revised variable. In addition, the

original specification fitted to the original sample may be wrong, merely reflecting the unique characteristics of that data set. Thus, one should not be surprised if replication using revised data, even for the same sample period, produces substantial changes in results.

Benefits of Confirmation

Even when replications are possible, they are done selectively because costs of replication exist. Thus, replications are reserved for experiments of unusual importance or for results that conflict with accepted theories or previous results (Committee on Conduct of Science, p. 11). Duplication of econometric results also is difficult, costly, and time consuming (see next section). Thus, it is important to elucidate the potential benefits of confirming published results.

Confirmation research may explain divergent econometric results. Divergent results arise from (1) differences in models, (2) differences in data, (3) uses of alternative estimators, including different computer codes, and (4) variations in the way results are analyzed and applied. Because so many reasons exist for differing results, confirmation is probably more important for analyses based on observational (nonexperimental) data than for data obtained from random samples or formal experiments. This is so precisely because the processes generating secondary data typically used by economists are largely unknown. Thus, applied economists usually pretest with a given data set to decide on a final model. The process of arriving at the final model is often neither well understood nor well explained. Indeed, the researcher may not be able to reconstruct how the final model was obtained, and different analysts may use quite different models to explain the same endogenous variable.

Consequently, one benefit relates to the fact that rival econometric models have proliferated and that these alternative models have similar statistical fits, though different economic interpretations (Hendry and Richard, p. 12). Models vary because they use different, possibly nonnested, concepts, but also because they contain different lag structures, functional forms, ways of measuring a concept, and sample periods. Thus, models vary not just because basic concepts are in dispute, but because analysts have treated details like lag structure and sample period differently. Analysts may agree, for example, that fed beef supply depends on expected prices, but disagree about whether rational or adaptive expectations is more appropriate. Even if they agree about the concept, the operational details for implementing the concept may differ. Different definitions of expected prices can result in striking variability in supply elasticities (Antonovitz and Green).

In my view, analysts should demonstrate that their results improve upon earlier work, stating explicitly what is meant by "improvement."[1] Confirmation of prior results is an important first step in establishing the relationship of one model to another. Casual comparisons of new with prior results are not sufficient; differences in results occur for many reasons, not just those of particular interest to the analyst. An author, for example, may attribute change to an improved definition of price risk when in fact the change is largely the result of an alternate sample. (Selection of the sample period is itself a modeling decision, and a change in variable definition may change the length of the sample.)

A second benefit, related to the first, is that the researcher learns from confirmation research. Obviously, the degree of learning will vary from case to case, but taking apart and putting together someone else's work can be illuminating. Thus, in my view, the potential for confirmation research to contribute to scholarly innovations has been undervalued. True scholarship arises from building in-depth expertise; confirmation provides greater understanding of the strengths and weaknesses of prior work; and it helps build the intellectual capital from which innovation can spring. Ladd, for example, points to the importance of imagination, analogy, metaphor, and simile in empirical research; thorough study of the work of others can contribute not only to synthesis, but to one's preparation for innovation.

Third, if a model is to be used for important decisions, it is essential that the results be as robust and as error-free as possible. Passell and Taylor, for example, approximately confirm results suggesting that capital punishment is a deterrent to murder, but in the process they also demonstrate that this result is highly sample specific. If the sample is modified slightly, the t ratio of the key coefficient changes from 3.0 to 0.4. Nevertheless, the original author makes

[1] Encompassing is one model-selection principle. According to Hendry and Richard (p. 17). ". . . one can ask of any specific model whether . . . it can account for the results obtained by other models; if so then the first model is said to be encompassing." Thus, ". . . encompassing is a central concept in any progressive research strategy . . . (p. 19)."

a spirited defense of the model. Whatever the appropriate model for explaining the murder rate, it is unquestionably true that many results are highly sensitive to changes in the model or sample; deletion of a single observation can cause significant changes in coefficients.[2,3] Obviously, it can be dangerous to establish public policy on the basis of uncertain results. (The fact that so few published studies have been confirmed has the unfortunate implication that the results have not been used for any serious purpose.)

Fourth, confirmation studies should encourage greater care and honesty in publication. Deliberate dishonesty in research probably is uncommon. But, few incentives exist for careful checking of data compilation and input, documentation of procedures (equivalent of keeping a lab notebook), and explanation of these procedures. Indeed, pressures exist for rushing to publication. Thus, I suspect that much careless work has been published. But, coming back to my first point, confirmation is not just an end in itself. It is a means to improve the quality of new empirical results.

Difficulties in Confirmation

Confirmation of published results requires duplication of the data set, models, and estimators. In addition, it may require duplication of applications of results, such as computing elasticities or simulations. None of these turn out to be easy.

Data

The principal reason for the difficulty in duplicating previous work is that the actual data used in the analysis are not available. Secondary data from governmental sources are subject to frequent revision, and many authors have not kept original data files.[4] Moreover, citations to data sources frequently are vague. Thus, the original data cannot be reconstructed.

The foregoing is compounded by three other problems. Errors can be made in data input; the actual sample period may differ from the stated period; and the nature of data transformations may not be clear. Even when computer files have been saved, problems can arise. Labels in the file may differ from those in the publication; the file contains original observations, but the transformations are unclear, or vice versa (e.g., Allen).

In considering the data problem, it is useful to think in terms of four different definitions of a particular concept: (1) the theoretical concept; (2) the true values of an observable variable which is used to measure the concept; (3) the observations actually available at time t on the observable variable; and (4) the data input by the researcher at time t. Many basic concepts in economics are not observable; they must be approximated by proxy variables. Moreover, the true values of the observable (proxy) variable are unknown; observed variables are often estimates subject to revision. Hopefully, the researcher has correctly entered the data available at time t (definition 3). Revisions presumably move one toward the true values of the observable variable (definition 2). Of course, the researcher may not actually use the data which are available at time t, either because of negligence in obtaining the most recent data or because of errors in data entry (definition 4).

Because of revisions, many vintages of a given data series exist. When entry errors are considered, the scope for differences in data sets is unlimited. If the original researcher has not kept the data file, it is virtually impossible to duplicate the research. Indeed, the original researcher probably cannot duplicate his or her own work in this situation.

Model

A second difficulty in confirmation, compounded by the first, is ambiguity about the ac-

[2] Heifner, commenting on Tomek and Gray, points out that one equation's coefficients change importantly if the first observation in the sample is deleted. Fortunately for Gray and me, the new result strengthened our argument, but it could have been otherwise.

[3] A subtly different issue arises from publication conventions and the effects of pretesting. If only "statistically significant" results are published, then a bias exists toward publishing the results of type I error because unpublished, nonsignificant results are unknown to other investigators, who therefore repeat the research independently (Sterling). With "enough" repetitions, a "significant" relationship will ultimately be obtained. The more likely scenario in economics is that a researcher is firmly committed to a particular hypothesis and continues to revise the model until the expected significant result is obtained. With pretesting, the probability of type I error grows enormously, and the result can easily be spurious (Wallace). If the research result is not confirmed and replicated, the spurious result is accepted as true.

[4] Time-series data from government agencies are revised frequently for a variety of reasons. For example, per capita pork consumption is computed from data on production, beginning and ending inventories, and net imports; on factors to convert from farm to retail weight; and on population estimates. All of these inputs into the series are subject to change, and since revisions occur at varying points in time, researchers attempting to reconstruct an historical analysis independently find it difficult to do so.

tual model fitted. The beginning and ending dates of variables may be unclear, especially when lagged variables are used. The precise specification of dummy (zero-one) variables may be uncertain, and if the coefficients are omitted from the published results, as is sometimes the case, it may not even be clear that the model contains dummy variables. Definitions of variables may involve weighted averages, ratios, products, or other transformations which are ill-defined and hence difficult to duplicate.

Sometimes models are fitted subject to restrictions, but the precise implementation of the restrictions is not made clear. Spline restrictions use specific knot locations, and published papers may be imprecise or in error about the actual locations used (e.g., Miller). Or, when imposing a polynomial lag structure, it may be unclear whether end-point restrictions have been used, and if so, precisely what the points are.

Computer Codes

A third difficulty in duplicating results relates to differences in computer codes. Generalized Least Squares, for example, is a generic estimator, and specific feasible GLS procedures vary from one econometric package to another. The difference between two procedures in correcting for first-order autocorrelation may be as simple as whether the first observation is retained. In Dewald, Thursby, and Anderson, results for a GLS estimator could not be confirmed even though the same data file as in the initial publication was used in the duplication attempt.

Computer programs also can contain errors. Early versions of SAS, for example, computed the Durbin-Watson statistic incorrectly. The limited literature on confirmation indicates that a variety of errors in privately written computer codes exist. (Dewald, Thursby, and Anderson mention two examples out of nine attempted confirmations.)

Potential problems of computational errors in confirmation research are complicated by the difficulties of using large data sets and/or complex estimators. Iterative estimation procedures can give different answers for a given data set because of differences in numerical methods, and it appears that identical computer programs can give different answers when installed on different computers. Dewald, Thursby, and Anderson failed to complete one replication attempt because of the costs inherent in transferring and using large data sets, and these costs arose even though the author of the original work was cooperative in providing the data.

Effect on Colleagues

A different type of cost in confirmation research is that it can be interpreted as a lack of trust in the integrity and competence of colleagues. This certainly is a potential issue in a relatively small sub-discipline like the econometric study of agricultural markets. We don't like to offend friends or colleagues. For example, an author cited in an earlier version of this paper protested and the citation has been deleted, even though the error cited was not very serious. As emphasized earlier, however, confirmation studies should be viewed primarily as a means of improving research, not as ends in themselves. Success in attempted confirmations is not the discovery of error by another analyst; rather, success should be measured mainly by the contribution of confirmation and replication to new and improved results.

Illustrations

In this section, I illustrate (1) the difficulty of duplicating published results and the related importance of obtaining the original data file, (2) the difficulties of using revised data, (3) the sensitivity of results to data revisions, and (4) the potential insights from attempted confirmations. The first example is drawn from Miller's unpublished MS thesis, which attempted to confirm selected studies of structural change in meat demand. Papers for confirmation were not selected because they were thought to be erroneous. Rather, criteria for selecting studies included the relative simplicity of the models and clarity of presentation.

One such study was a paper by Braschler, which fitted traditional price-dependent demand equations to annual observations. Using the sample period 1950–82, Braschler concluded, among other things, that a structural break occurred in the demand for beef between 1970 and 1971; i.e., the data partition 1950–70, 1971–82 minimized the total sum of squared errors. The corresponding F statistic was 11.12.

Miller first made an independent attempt to confirm Braschler's results. The resulting coefficients are close, but not identical, to those published. Fortunately, Professor Braschler had saved his data files, and with these files it was

possible to exactly confirm the published coefficients (table 1). Miller's independently collected income data were somewhat smaller values than those reported by Braschler for the years 1970–82; a few other data points had small differences; ambiguity also existed about whether the author shifted to the CPI-U from the CPI-W series when CPI-U became available in 1978.

Having confirmed the published results, Miller attempted a replication with revised data. Observations on all variables in the equations had been revised by government agencies, most more than once. One problem was that a consistent set of revisions was unavailable for each variable over the entire sample period. For example, the latest (as of 1991) revisions for beef, pork, and chicken consumption were available only from 1955 onward.

Two approaches to this problem are possible. One is to join the data end-to-end without adjustment, and this is probably the approach used by most analysts when they combine revised with unrevised series. A second approach is to adjust the older data to obtain a more consistent series. Specifically, Miller employed the new (revised) observations as the dependent variable in a regression model in which the old observations form the explanatory variable. Additional regressors, like a linear trend, were used in the "data adjustment" model, if judged appropriate. The fit from the overlap period is then used to make backward forecasts for the years not covered by the revisions. These forecasts are the estimated revisions. In table 1, the column "revised data (1)" presents the results from joining old and new data end-to-end; the column "revised data (2)" presents the results using the adjustment process.

A second potential problem in using revisions is changes in units of measure or scale. In the example, the CPI shifted to a more recent base period. This makes the revised CPI smaller; consequently, the deflated price (dependent variable) is larger when the revised CPI is used. Thus, even if no other change occurred, the estimated coefficients would be larger. A variety of approaches to this problem is possible; one is to shift the base period of the new index back to the older period. This would have had the consequence of changing the beef consumption coefficient in data set (1) from −2.03 to −0.68, which compares with −0.84 in the original research. It appears that the effect of revisions and of adding new (hence revised) data points is to reduce the absolute magnitude of the slope coefficient of beef consumption (Miller).

The emphasis of the Braschler paper, however, is on structural change, and the scale problem is irrelevant to the F test. Miller's analysis demonstrates that data revision alone shifts the structural change dates. With data set (1), the structure is estimated to change between 1972

Table 1. Inverse Demand Functions, Beef, U.S., 1950–82

			Revised data[c]		
Variables[a]	Reported and confirmed	Independent attempt	(1)	(2)	(3)
Intercept	80.25	79.01	214.5	267.3	176.3
	(5.33)[b]	(4.17)	(3.65)	(2.86)	(2.14)
QBF	−0.836	−0.814	−2.093	−2.030	−1.649
	(7.00)	(6.02)	(6.00)	(3.95)	(2.81)
QPK	0.165	0.091	0.346	−0.040	0.797
	(0.92)	(0.40)	(0.69)	(0.05)	(1.11)
QCH	−0.647	−0.753	−3.161	−3.770	−7.138
	(2.57)	(2.72)	(3.26)	(3.41)	(5.43)
INC	0.037	0.040	0.041	0.041	0.052
	(5.43)	(5.08)	(4.96)	(4.038)	(4.40)
R^2	0.716	0.677	0.615	0.548	0.686
d	0.99	0.99	0.92	1.25	1.11

[a] Dependent variable is retail price beef, deflated by CPI, cents per lb.; QBF, QPK, QCH are consumption of beef, pork, and broilers, lb. per capita; INC is disposable personal income, deflated by CPI, $ per capita; *d* is the Durbin–Watson statistic.
[b] t-ratios given in parentheses
[c] (1) and (2) represent two ways of using revised data, see text. Both shift CPI base from 1967 = 1.0 to 1982–4 = 1.0 and use 1950–82 sample. Column (3) uses a 1958–90 sample, but is otherwise consistent with column (2).
Source: Miller

and 1973, while with data set (2) the structure is estimated to change between 1958 and 1959. Thus, not only does the estimated date of structural change shift, the date depends on how the revisions are joined to the early unrevised data. Also, both equations using revised data have smaller R^2's than in the original sample, and the residuals appear to be autocorrelated in all data sets. Thus, the F tests are suspect.

Using revised data for the original sample period changed the conclusions of the analysis. If the sample period is shifted to more recent years, results change again (table 1), and if the model is modified, still additional changes occur (Miller). The important point is that the use of revised data for the original time period changes Braschler's result; the original paper could be duplicated, but the result is not replicated with revised data (for a critique of analyses of structural change, see Alston and Chalfant). This analysis then becomes the base for analyzing the effects of a new sample and/or a new model.

My second example is based on a paper suggesting that the U.S. fed beef sector might best be modeled as in disequilibrium (Ziemer and White). Because disequilibrium models had almost no applications in agricultural economics, those interested in the economics of the fed beef sector—or agricultural product markets in general—could potentially benefit from a careful analysis of Ziemer and White's paper. It contained a novel idea, and at that time it was well-worth the effort to understand the results and how the model worked. Such analysis could possibly suggest improved modeling strategies.

As it turned out, Shonkwiler and Spreen could not independently confirm the results. The coefficients in the demand equation, based on data sources used by Shonkwiler and Spreen, were especially different from those reported in the original article (table 2). However, Ziemer and White had kept the data file, and Ferguson was able to duplicate exactly the original results using the authors' data set. Unlike most confirmation studies, duplication of the original results was relatively easy.

Ferguson, however, was not able to exactly match the income series in Ziemer and White's computer file with any published series on per capita personal income. Moreover, two income observations in the computer file were approximately 20% too large, both from the perspective of the immediately preceding and following data points in the file and from the perspective of similar published series, and were clearly in error. When the two observations were corrected, the coefficients estimated by Ferguson roughly approximated those obtained by Shonkwiler and Spreen with independently collected data (table 2).

Clearly, without the assistance of the authors, it would not have been possible to duplicate Ziemer and White's published model. The original data series could not be independently duplica-

Table 2. Fed Beef Demand, Farm Level, U.S., 1965–79

Variables[a]	Reported and confirmed	Confirmation attempt[c]	Modified confirmation[d]
Intercept	588.0	-4708	-.3888
	(0.8)[b]	(4.70)	(4.99)
Price fed beef	-72.01	157.8	-132.1
	(1.78)	(4.50)	(4.26)
Price utility cows	89.75	147.8	119.8
	(2.32)	(4.65)	(4.18)
Price hogs	-9.15	16.94	-12.84
	(0.64)	(1.50)	(1.24)
Real income	172.2	455.8	344.8
	(6.38)	(10.16)	(11.60)
Durbin–Watson	—[e]	--	0.66

[a] Dependent variable is marketings of fed cattle (see Ziemer and White).
[b] Coefficients in parentheses are ratios of coefficients to standard deviations; equations estimated by two-stage least squares.
[c] Confirmation attempt by Shonkwiler and Spreen uses independently collected data. Coefficients reported here are for the Ziemer and White model. In addition, Shonkwiler and Spreen modify the original model, thereby magnifying the differences between those reported by Ziemer and White and those in Shonkwiler and Spreen.
[d] Ferguson confirmed original result with authors' data file; her research corrects income series (see text), but otherwise retains original data file.
[e] Not reported

ted, but by using the authors' file with the two points corrected, it was possible to explore alternative models and estimators and to compare them with the original results (Ferguson). In this case, the residuals of the demand equation remained autocorrelated after the data errors were corrected, and the reasons for this autocorrelation needed examination. Further analysis suggested that the fed beef market was probably not in disequilibrium, at least in the sample period studied.

Published and anecdotal evidence on confirmation in economics suggests the disheartening conclusion that many published empirical studies contain errors and that some of these errors are serious in the sense that, if corrected, the stated conclusions of the study would change.[5] Those of us who have published empirical results probably have published errors, and perhaps drawn incorrect conclusions. Thus, while I have stressed the insights developed from confirmation studies, one cannot ignore the need to correct errors, when necessary, and to discourage careless analysis and writing.

Incentives for Improved Scholarship

The highest rewards in science are reserved for innovations, and research time is scarce. Moreover, empirical research in economics is often complex, and hence confirmation is potentially time-consuming. The details of a research project are extremely difficult to reconstruct if the original researcher has not followed a protocol that preserves the data and modeling procedures. Yet, it is precisely such details that often are left to support staff and/or graduate students and that are often lost when projects and theses are completed. Wible points out that replication failure is understandable in economic terms: a researcher's overriding objective is to maximize his or her own expected utility, and this utility is maximized by the prompt publication of (hopefully) "innovative research." This discourages confirmation research on the one hand, and on the other does not provide payoffs for maintaining records which can be used by those interested in confirmation.

If individual researchers do not have incentives to attempt to confirm published research, but from the profession's viewpoint benefits of confirmation exceed costs, a type of market failure is occurring. Incentives would change, however, if unconfirmable research is seen as an inferior good and if confirmation is viewed as contributing to true scholarship. These precisely are the messages of the present paper. Confirmation helps build the intellectual capital for innovation, and duplication provides the basis for determining whether a model is adequate and whether new results improve upon the old (i.e., that the new results really are innovative).

Of course, confirmation studies should be done selectively. Not every published paper deserves in-depth appraisal. In research, literature reviews identify the intellectual bases for new work, and such reviews suggest key models and results. It is the key models that deserve attempts to confirm and replicate. Thus, such analysis becomes a part of an appraisal of the literature, and in this context, confirmation and replication attempts are an implied compliment to the earlier work.

Costs of confirmation, however, must be reduced. In principle, duplication of previous work using time-series observations should be less expensive than repeating an experiment. But researchers must keep data files, special computer programs, and the details of research methods. It is now possible to save data inexpensively in electronic form. Nonetheless, it is still time-consuming to label files carefully and to maintain precise records of research procedures. The latter requires a change in attitude, which cannot be achieved by exhortation. Rather, failure to keep good records must be made more expensive. Results that are not carefully documented must be treated as inferior.

Thus, like Dewald, Thursby, and Anderson, I urge that professional journals require authors to submit programs and data. The paper itself or an appendix to it (which may or may not be published) should contain a full description of the data and data sources, transformations, model restrictions, and estimator. These descriptions must be precise and complete, including such details as the dates covered by each variable (especially when lags are involved). In addition, authors must make a good-faith effort to remove ambiguities in their writing. Small details, like

[5] The confirmation literature supports these conclusions. Dewald, Thursby, and Anderson were only able to confirm two out of nine articles in their entirety. They also reproduced most of the results of a third article and obtained qualitatively similar results for a fourth. They noted a number of errors as they attempted their confirmations. In an exhaustive analysis of two articles, Miller duplicated most of the results of one article, but could not confirm the results of a second even with the assistance of an author. The Ziemer and White results are based on two data entry errors. Leimer and Lesney's critique of Feldstein notes an important error reversing a sign and hence the conclusions of his study (also see Allen; Passell and Taylor).

publishing means of the variables, are helpful and take little space. Papers should be written so that a careful reader can, in principle, duplicate the results.

One question a paper's reviewers should ask is, could I duplicate the quantitative results in this paper if asked to do so? In the not-too-distant future, improvements in technology should make duplication relatively cheap, and perhaps new standards will require duplication of results prior to publication, including tests for the robustness of results.

If new empirical results truly build on prior work, authors must demonstrate that the new results improve upon previously published results. Such a demonstration must be more than a casual comparison of coefficients. Both the old and new model should be fitted to the original data file (used for the old model) and to new data now available. Does the new model really improve upon the old with both data sets? That is, the new model should encompass the old, and the new model should be robust over different sample periods; or if it is not, the changes in coefficients should be explainable. New results should increase our understanding, not add to existing confusion.

In sum, agricultural economics (and applied economists in general) must set higher standards of excellence in empirical research. Higher-quality output requires both more and better inputs in terms of model specification, data, and the researcher's intellectual input. It is my view that adding confirmation as an initial component of the research agenda will improve quality. Admittedly, it will lengthen the research process and probably result in fewer published papers, but the profession profoundly needs to establish higher standards for published empirical results. We need a radical change in the way empirical research is conducted. This will occur only if we develop incentives to achieve improved levels of scholarship.

References

Allen, D. W. "Marriage and Divorce: Comment." *Amer. Econ. Rev.* 82(June 1992):679–85.

Alston, J. M., and J. A. Chalfant. "Can We Take the Con Out of Meat Demand Studies?" *West. J. Agr. Econ.* 16(July 1991):36–48.

Antonovitz, F., and R. Green. "Alternative Estimates of Fed Beef Supply Response to Risk." *Amer. J. Agr. Econ.* 72 (May 1990):475–87.

Bonnen, J. T. "Improving the Socioeconomic Data Base." *Agriculture and Rural Areas Approaching the Twenty-First Century*, ed. (chapter 15) R. J. Hildreth et al. Iowa State University Press, 1988.

——— "On the Role of Data and Measurement in Agricultural Economics Research." *J. Agr. Econ. Res.* Supplement (July 1991):15–18.

Braschler, C. "The Changing Demand Structure for Pork and Beef in the 1970s: Implications for the 1980s." *South. J. Agr. Econ.* 15(December 1983):105–10.

Committee on the Conduct of Science, National Academy of Sciences. *On Being a Scientist*. Washington, D.C.: National Academy Press, 1989.

Dewald, W. G., T. G. Thursby, and R. G. Anderson. "Replication in Empirical Economics: The Journal of Money, Credit and Banking Project." *Amer. Econ. Rev.* 76(September 1986):587–603.

Ferguson, C. A. *An Evaluation of a Disequilibrium Model*. Cornell Univ. A. E. Res. 83–17, August 1983.

Heifner, R. G. "Temporal Relationships Among Futures Prices: Comment." *Amer. J. Agr. Econ.* 53(May 1971):361–62.

Hendry, D. F., and J.-F. Richard. "On the Formulation of Empirical Models in Dynamic Econometrics." *J. Econometrics* 20(October 1982):3–33.

Ladd, G. W. "Artistic Research Tools for Scientific Minds." *Amer. J. Agr. Econ.* 61(February 1979):1–11.

Leamer, E. E. "Let's Take the Con out of Econometrics." *Amer. Econ. Rev.* 73(March 1983):31–43.

Leimer, D. R., and S. D. Lesney. "Social Security and Private Savings: New Time-Series Evidence." *J. Polit. Econ.* 90(June 1982):606–29.

Miller, D. J. *Assessing Studies of Structural Change in Meat Demand*. Cornell University, MS Thesis, 1991.

Mittelstaedt, R. A., and T. S. Zorn. "Econometric Replication: Lessons from the Experimental Sciences." *Quart. J. Bus. and Econ.* 23(Winter 1984):9–15.

Passell, P., and J. B. Taylor. "The Deterrent Effect of Capital Punishment: Another View." *Amer. Econ. Rev.* 76(June 1977):445–51.

Patterson, K. D., and S. M. Heravi. "Efficient Forecasts or Measurement Errors: Some Evidence for Revisions to United Kingdom GDP Growth Rates." *The Manchester School of Economic and Social Studies* 60(September 1992):249–63.

Shonkwiler, J. S., and T. H. Spreen. "Disequilibrium Analysis: An Application to the U.S. Fed Beef Sector: Comment." *Amer. J. Agr. Econ.* 65(May 1983):360–63.

Spanos, A. *Statistical Foundations of Econometric Modeling*. Cambridge: Cambridge University Press, 1986.

Sterling, T. D. "Publication Decisions and Their Possible Effects on Inferences Drawn from Tests of Significance—or Vice Versa." *J. Amer Statist. Assoc.* 54(March 1959):30–34.

Wallace, T. D. "Pretest Estimation in Regression: A Survey." *Amer. J. Agr. Econ.* 59(August 1977):431–43.

Wible, J. R. "Maximization, Replication, and the Economic Rationality of Positive Economic Sciences." *Rev. Polit. Econ.* 3(April 1991):164–86.

Ziemer, R. F., and F. C. White. "Disequilibrium Market Analysis: An Application to the U.S. Fed Beef Sector." *Amer. J. Agr. Econ.* 64(February 1982):56–62.