MARKETING SCIENCE
Vol. 25, No. 3, May–June 2006, pp. 203–216
ISSN 0732-2399 | EISSN 1526-548X | 06 | 2503 | 0203

DOI 10.1287/mksc.1060.0215
© 2006 INFORMS

Editorial: Errors in the Variables, Unobserved Heterogeneity, and Other Ways of Hiding Statistical Error

Steven M. Shugan*
University of Florida, Warrington College of Business, 201B Bryan Hall, P.O. Box 117155,
Gainesville, Florida 32611, steven.shugan@cba.ufl.edu

One research function is proposing new scientific theories; another is testing the falsifiable predictions of those theories. Eventually, sufficient observations reveal valid predictions. For the impatient, behold statistical methods, which attribute inconsistent predictions to either faulty data (e.g., measurement error) or faulty theories.

Testing theories, however, differs from estimating unknown parameters in known relationships. When testing theories, it is sufficiently dangerous to cure inconsistencies by adding observed explanatory variables (i.e., beyond the theory), let alone unobserved explanatory variables. Adding ad hoc explanatory variables mimics experimental controls when experiments are impractical. Assuming unobservable variables is different, partly because realizations of unobserved variables are unavailable for validating estimates.

When different statistical assumptions about error produce dramatically different conclusions, we should doubt the theory, the data, or both. Theory tests should be insensitive to assumptions about error, particularly adjustments for error from unobserved variables. These adjustments can fallaciously inflate support for wrong theories, partly by implicitly under-weighting observations inconsistent with the theory. Inconsistent estimates often convey an important message—the data are inconsistent with the theory! Although adjustments for unobserved variables and ex post information are extraordinarily useful when estimating known relationships, when testing theories, requiring researchers to make these adjustments is inappropriate.

Key words: unobserved heterogeneity; scientific method; falsification; statistical validation; errors in the variables; Popper falsification

Theories and Observations

This section argues, from a theory-testing viewpoint, that the primary role of assuming measurement error is to save theories from falsification by weakening the tests required by the scientific method. We begin by reviewing the definition of a scientific theory and the scientific method for testing theories.

In many ways, theories are potential laws that still have disbelievers. Laws and theories predict that particular events will occur while others will not. Scientific theories, unlike pseudo-science, make falsifiable deterministic predictions (Popper 1965, pp. 33–37). To be empirically falsifiable, the theory must divide events into two classes: those consistent and those inconsistent with the theory (Popper 1968, p. 86). For example, common predictions from physics include: (1) Bodies continue in their state of constant velocity unless acted upon by an external force; (2) bodies attract each other with equal and opposite forces; and (3) the total amount of energy in a closed system will remain constant. Often, discovering theories is far more difficult than testing them. For example, discovering that mosquitoes spread malaria by carrying parasites is more difficult than testing the theory. General, replicable, falsifiable, and valid theories are the goal of science. When testing theories, we must use great care when we assume ad hoc error terms, ad hoc variables (observed or unobserved), and specific forms of randomness (e.g., distributions, covariance, etc.) that are all absent from the theory itself in order to fix defects in the data. Ideal theory tests should only include the variables in the theory.

All scientific theories must be falsifiable (Popper 1972, pp. 47–48). Then, employing the scientific method, we reject theories making false predictions. For example, one theory (Syam et al. 2005) predicts that, in equilibrium situations, when firms can

Editorial pages are not part of the regular *Marketing Science* page budget. We thank the INFORMS Society of Marketing Science for paying for all editorial pages. We also thank the Society for granting every page supplement requested by the current editor.

We welcome and often post responses to editorials. Please see mktsci.pubs.informs.org.

*Steven M. Shugan is the Russell Berrie Foundation Eminent Scholar in Marketing.

customize on more than one attribute, we should not observe full customization but we should observe either partial or no customization. We might reject the theory if we observe full customization. Another theory (Wu et al. 2004) predicts that, when provision of consumer information service is costly and consumers can easily free ride on that information, we should not observe sellers who free ride all the time. We might reject the theory if we observe full free riding. Still another theory (Liu and Zhang 2006) predicts that when retailers can implement personalized pricing and the manufacturer can leverage both personalized pricing and entry into a direct distribution channel, retailers would only adopt personalized pricing to deter entry and not for profitability. We might reject the theory if we observe retailers adopting personalized pricing to increase profitability.

Moreover, if a theory fails to provide sufficiently accurate definitions for the theory's terms, we might reject the theory as ambiguous. If a theory fails to make testable implications, we might reject it as tautological. If the theory's definitions are insufficiently precise to make testing doable, we might reject the theory as inchoate. If the theory makes predictions that are conceptually unobserved, we might reject the theory as unverifiable. Of course, with assiduous improvement or discovery of new observations, we might resurrect rejected theories. For example, Keuzenkamp and Barten (1995) argue that tests of Marshallian demand functions lead to a history of rejections without falsification as economists searched for a better specification of the theory. We might also define new boundary conditions to limit claims about when the theory is valid. For example, female mosquitoes spread malaria only after biting someone with malaria. The scientific approach warns against overconfidence in the theory, which often leads to problematic outcomes (e.g., see Montgomery and Bradlow 1999).

Some theories inexplicably endure despite falsification as proponents struggle with derivative theories. For example, Keynesian theory enjoys some popularity despite observation of events in the 1970s that were manifestly inconsistent with the theory, i.e., simultaneous unemployment and inflation (Weintraub 1999). Marxist-Leninist theory faced a similar situation with inconsistencies between events and theoretical prediction observed in the Soviet bloc. Perhaps economic historian Mark Blaug (1980) was correct when he claimed economists practice innocuous falsification. In contrast, Redman (1994) argues that the traditional scientific method fails to apply to economics, in part because we can find observations that contradict all economic theories and economic truths. That is a sobering thought.

Although falsification might be far too severe a standard for most budding theories and struggling academics facing tenure decisions, validation in the form of testable predictions (i.e., predictions that can reject if not falsify) is critical. Kleindorfer et al. (1998) present many weaker criteria for validation (e.g., professional acceptance, having a rational foundation, predominance of empirical success). Friedman (1953), in his famous defense of economic models of consumer behavior, argued that models are always only unrealistic abstractions and their evaluation depends not on the realism of their assumptions but on the accuracy of their predictions. A good theory yields valid and meaningful predictions. However, all of these viewpoints do agree that sufficient inconsistencies between theory and data should lead to serious doubt, if not complete rejection, of the suspect theory. Theory falsification must be possible, at least within certain boundary conditions (e.g., the theory only works for durables). If an empirical application is unable to falsify a theory, the application might be useful, but it is not a test of the theory.

The usefulness of the theory often lies in whether it can make predictions that we could not make without the theory (Shugan 2005) and the ease with which we can observe the relevant conditions and predictions. For example, a theory that predicts unobserved outcomes from unobserved conditions provides little or no current value. Perhaps the ultimate goal of theories should be errorless verifiable predictions that no other theory has previously made.

Hence, theories make predictions, and we need only observe those predictions to test the theory. We might ask whether statistics plays any role in theory testing. The following discussion argues that some statistical methods help us save theories that are inconsistent with observations. These methods can help us find excuses (often legitimate) for why our favorite theory is inconsistent with observation. The most common excuse is measurement error. Although we expect that advances in measurement technology will eventually validate our theory,[1] for the impatient, we have statistics. Statistics helps us interpret whether defects in the data are causing the inconsistencies between our theory and observation. However, as Healy (1978, p. 385) states: "Statistics may have a more important role to play in technology than in science; it may itself best be considered as a technology rather than a science." Of course, statistics remains an invaluable tool in marketing.

Parenthetically, usually scientific research produces multiple competing theories that are all perfectly consistent with all available observations at the time

[1] If we are unable to validate a theory with ideal measurement, it appears our theory is a tautology.

(e.g., classical Newtonian physics versus the theory of relativity, the big bang theory versus the static universe). Many political debates are similar. Each side often attributes observed outcomes to different causes and, hence, each side advocates competing theories, each fitting all currently available observations. As John D. C. Little once remarked at a keynote address, great advances in knowledge usually follow advances in data collection. One reason is that new data allow discrimination between multiple theories that all perfectly predict the old data. Without discriminating empirical theoretical tests, we might use parsimony (i.e., Occam's razor) as the relevant criteria (e.g., see Shugan 2002). Eventually, advances in measurement technology should solve the problem. For example, testing theories of curved-space required advances in telescope technology (van Helden 1974). The Heisenberg uncertainty principle does dictate a limit for the accurate measurement for subatomic particles.[2] In most other cases, however, we are often far from that limit in the social sciences.

In sum, attempting falsification is the critical step in testing scientific theories. Assuming measurement error (and other errors discussed in later sections) tends to make falsification more difficult in the sense that we forgive some inconsistencies between theoretical predictions and observation. Some probabilistic predictions might be impervious to strict falsification (Gillies 1971) without imposing additional conditions.

A Note on Notation and Heterogeneity

Subsequent sections use the same basic notation. Theories predict some events and not others. For example, consider the theory that marketing activities have threshold effects. We should observe s-shaped advertising response functions (Little 1979, Vakratsas et al. 2004). We should observe turning points in new product life cycles (Golder and Tellis 2004). We should observe pulsing (Mahajan and Muller 1986, Villas-Boas 1993). We should not observe responses to low levels of marketing activity. Observing some patterns of advertising by experienced advertisers would refute the theory (or require a more restrictive statement of the theory).

Most theories predict the occurrence of joint events (x_j, y_j) and not other events $(\overline{x_j, y_j})$. We can sometimes translate that joint prediction into a true conditional relationship $y_j = \beta_0 + \beta_1 x_j$ between variables x_j and y_j for every observation j. Both variables vary as a pair across observations j. The constants β_0, β_1 are model parameters where $\hat{\beta}_1$ denotes an estimate for β_1. The random variables $\tilde{x}_j$, $\tilde{y}_j$ represent the observed x_j, y_j. Random variables $\tilde{\varepsilon}_j$, $\tilde{w}_j$ are independent and with zero mean so that, for example, $E[\tilde{\varepsilon}_j \tilde{w}_j] = \sigma_{\tilde{\varepsilon}_j \tilde{w}_j} = 0$ and $E[\tilde{\varepsilon}_j^2] = \sigma^2_{\varepsilon_j}$ where $E[\cdot]$ denotes the expectation operator, $\text{var}[\varepsilon_j] = \sigma^2_{\varepsilon_j}$ denotes the variance of variable $\tilde{\varepsilon}_j$, and $\text{cov}[\tilde{\varepsilon}_j, \tilde{w}_j] = \sigma_{\tilde{\varepsilon}_j \tilde{w}_j}$ denotes the covariance of variables $\tilde{\varepsilon}_j$, $\tilde{w}_j$.

When the observed value of $\tilde{x}_j$ differs from the true x_j, we could attribute some of that difference to heterogeneity rather than to measurement error. For example, two individuals $j = 1, 2$ may appear to be the same $\tilde{x}_1 = \tilde{x}_2$ but are different $x_1 \neq x_2$. Hence, a given x_j maps to a random variable $\tilde{x}_j$.

From this point forward, for simplicity in presentation, each section examines a different model for $\tilde{y}_j$ and, therefore, the mathematical definition of $\tilde{y}_j$ varies by section. Finally, for notational simplicity, subsequent sections drop the redundant subscript j until we consider multiple observations for each individual. Expectations are over j.

Most Theories Have No Dependent Variables and No Error

This section argues that critical step in translating a theory into a statistical model is the decision regarding where to include the error term. That decision, for example, determines which variables are dependent and which are independent.

Consider the translation of a theory into a statistical format. That format often (not always) includes the statistical concepts of dependent (i.e., y) variables, independent variables (i.e., x), and statistical error. Unfortunately, these statistical concepts are absent from most theories. Moreover, where to include statistical error and the designation of variables as either dependent or independent is arbitrary in the sense that there is usually no guidance from the theory itself.[3]

Often, we make the distinction between dependent and independent variables based on where we want to include statistical error. If we can measure variables without error, we call them independent variables. Variables with error become dependent variables.

Causality is irrelevant to this distinction. There are some theories that predict causal relationships. For example, Zoltners and Sinha (2005) argue that poor territory alignment causes undesirable variation in incentive payouts, making compensation plans appear wrong. Dubé and Manchanda (2005) argue

[2] Eric T. Bradlow (personal communication) argues that there might be inherent randomness in consumer responses.

[3] Some theories include randomness but often derived from a deterministic idea. For example, quantum mechanics has randomness, but the foundation is the deterministic idea of Heisenberg. Gambling has randomness, but the foundation is the deterministic idea that casinos are trying to create unpredictable events (they sometimes fail). This discussion only claims that most theories do not include error.

that lower media availability in smaller markets causes more competitive advertising (share stealing) than in larger markets where high media availability causes complementarity (category advertising). Kalnins (2004) argues that incentives, from the governance form of franchising rather than simply the outcome of expansion, cause encroaching behavior.

Although causality is often the prime motivation for referring to y as the predicted or dependent variable, there is no element of causality in statistical analysis, which only relies on correlation. The belief that there is necessarily a causal relationship between dependent and independent is wrong. Some textbooks exacerbate the problem by referring to structural models as causal models. Statistical tests examine correlation, and only theory infers causality. Kluge (2001), for example, explains that likelihood functions are unable to decipher causality.

There have been attempts to use observed event sequences to evidence causality in statistics (e.g., the Granger-causality test). Of course, sequencing is no guarantee of causality. We can observe the effect before we observe the cause. Politicians (sometimes with greater current vote totals) concede elections before final vote totals are known, but future vote totals are the cause. We observe increased retailer orders shortly before Christmas, but Christmas causes these orders. We observe tides before seeing the moon, yet the moon causes the tide. We hear a train whistle before seeing the train, yet the train causes the whistle. In fact, we observe phone calls to 911 emergency before observing people in distress, but 911 emergency calls do not cause the distress. We observe symptoms of bedbugs before observing the bug (or the bite, for that matter), because bedbug saliva contains anesthetic to numb the pain. Moreover, causality might have no observable temporal sequence. For many people, alcoholic drinking causes depression and, at the same time, depression causes alcoholic drinking.

Without experimental data, we can link directed acyclic graphs, corresponding to factorizations of the joint probability distribution, to causality (Pearl 2000); however, we still require theory to dictate the direction of possible causality. With experimental data, we have much greater evidence of causality, but we still require theory to disentangle confounds, construct manipulation checks, eliminate demand artifacts, and construct an appropriate design.

In sum, causality is not the reason to categorize variables as dependent or independent[4]—statistical error is.

What Is Statistical Error?

This section argues that statistical error is simply the inconsistency between the proposed theory and observation.

We can translate the predictions of some theories into a linear equation, i.e., $y = \beta_0 + \beta_1 x$ with parameters β_0 and β_1, where y is the dependent variable and x is the independent variable. Suppose the theory predicts this form and, possibly, β_0 and β_1 as well.

As noted earlier, the decision regarding which variable to predict is often a decision about where to include statistical error. Statistical error goes by myriad names, including random error, measurement error, structural error, residual error, sampling error, disturbances or, simply, error. Failing to observe y conditional on x should cast doubt on the theory because the absence of correlation is evidence against the theory. However, hoping to rehabilitate our theory, we might argue that, if not for measurement error when observing y, we would have confirmed our theory's predictions. Hence, the linear model, given in Equation (1), is born. Of course, there are many examples of when linear models are only building blocks for nonlinear models (e.g., Evgeniou et al. 2005).

$$\tilde{y} = \beta_0 + \beta_1 x + \tilde{\varepsilon} \qquad (1)$$

We might argue that the observed $\tilde{y}$ differs from the true y by no fault of our theory. The random measurement error $\tilde{\varepsilon}$ absolves our theory from making perfect predictions. Of course, adding $\tilde{\varepsilon}$ should provide no better prediction of y, given x because $\tilde{\varepsilon}$ is random. Being random is convenient for saving our theory because, by definition, random often implies unpredictable.[5] Hence, it is, by definition, futile to remove random error completely and verify our theoretical prediction with certainty.

When the absolute error $|\tilde{\varepsilon}|$ is sufficiently large, we must abandon our excuses and admit that the data cast doubt on our theory. In that event, we might refer to $|\tilde{\varepsilon}|$ as structural error or error in the theory, rather than as merely innocuous (if not helpful) measurement error. We could proudly refer to this procedure as accounting for random error. However, that statement borders on being disingenuous, because accounting for random error favors confirmation of our theory over rejecting it.

There is absolutely no intent here to be critical of the simple linear model. This extraordinarily powerful concept has obviously lead to remarkable

[4] It is unclear whether we could design and implement an experiment to demonstrate causality with no theory. For example, it might appear that turning off a television causes television broadcasting to stop. In 1632, Galileo refuted Aristotle's causal theory of gravity (Wisan 1984) with a simple thought experiment (i.e., tying two stones together). In fact, improperly disentangling causality has been the devastating undoing of most past theories now in disrepute.

[5] It might be predictable when conditioned on other observations or other information.

advances in knowledge. Moreover, some attempt to infer causality is better than no attempt. However, we should avoid the pretence that accounting for random error somehow elevates our standards for confirming theories. In contrast, accounting for random error often favors our theory because it ascribes some prediction inconsistencies to faulty data rather than to faulty theory. We should admire theories that make unconditional absolute predictions with no excuses. In sum, error is always bad—regardless of the name ascribed to it. As Popper states: "although we seek theories with a high degree of corroboration, *as scientists we do not seek highly probably theories but explanations; that is to say, powerful and improbable theories*" (Kluge 2001, p. 322, italics in original).

Errors in the Variables as a Way to Hide Error

This section argues that all errors reflect defects in the theory, regardless of why or where we include them in the model. When different statistical assumptions about error produce dramatically different conclusions when theory testing, we must doubt the theory, the data, or both. Theory tests should not be extremely sensitive to statistical assumptions (about error or required adjustments for defects in the data).

Consider the apparently more complex situation when we assume random error in the independent variable x. Outside of an experimental setting, we only observe x, and we do not necessarily control it. Consequently, we introduce another error $\tilde{w}$. We define $\tilde{x}$ as the observed x, where $\tilde{x} = x + \tilde{w}$ suffers from contamination by measurement error.

In this case, our designation of x as an independent variable seems completely arbitrary because both x and y contain statistical error. Hence, we have hopelessly blurred the distinction between dependent and independent variables, because both variables now contain error. From a testing viewpoint, the question is whether we observe the theoretically predicted joint event (x, y) and not observe events inconsistent with the theory.

A cynical critic might view $\tilde{w}$ as still another excuse for inconsistencies between theoretical predictions and observations. It is. However, if errors contaminate observations y, the natural conclusion is that errors contaminate the observations of x.

Still another justification for a random independent variable $\tilde{x}$ is heterogeneity. It is logical to assume that if observations vary across individuals, the error $\tilde{w}$ might capture unobserved individual variation or heterogeneity. More on that later.

In any case, we should view both $\tilde{e}$ and $\tilde{w}$ as bad. They are not different in nature. Both reflect departures from theoretical predictions. Large absolute errors are always bad and place our theory in jeopardy because observations fail to conform to theoretical predictions. However, we sometimes view $\tilde{w}$ differently than $\tilde{e}$. This practice is very dangerous for the following reason. Consider the modified simple linear model in Equation (2).

$$\tilde{y} = \beta_0 + \beta_1(x + \tilde{w}) + \tilde{e} \tag{2}$$

Equation (2) explicitly assumes (allows?) independent variable error, i.e., $\tilde{x} = x + \tilde{w}$. Rearranging Equation (2) yields Equation (3).

$$\tilde{y} = \beta_0 + \beta_1 x + \beta_1 \tilde{w} + \tilde{e} \tag{3}$$

Note that even if we knew the parameters β_0, β_1 there would still be an infinite number of values for $\tilde{e}$ and $\tilde{w}$ that produces exactly the same total error $\beta_1\tilde{w} + \tilde{e}$ that describes the deviation between the observed $\tilde{y}$ and the y predicted by our theory.

It appears that we have partitioned the error into two components. The first is a good error $\beta_1\tilde{w}$ that reflects innocuous measurement error and, possibly, heterogeneity. The second is a bad component $\tilde{e}$ that reflects possible structural error in our theory.

This is obviously an illusion because, in the end, whether we blame total prediction errors on either the theory or the data, the total error $\beta_1\tilde{w} + \tilde{e}$ provides exactly the same evidence for or against the theory. The addition of ad hoc unobservable variables rarely improves the prediction (versus fit) of a truly correct theory because unobserved variables conceptually provide no information to enhance predictions (i.e., unless the theory has defects). Our best hope for improving the ability to predict is to add new information, e.g., additional observable variables, or to adopt a new theory. Note that using out-of-sample information about the distribution of $\tilde{w}$ does qualify as new information; however, as discussed later, there is some danger involved when the goal is testing theories. Shrinkage techniques, for example, might rightfully adjust $|\beta_1|$ or the standardized $|\beta_1|$ to reflect out-of-sample information (e.g., the average $|\beta_1|$ estimated for other individuals). Those techniques often provide better estimates. When testing theories, however, the scientific method warns against employing (explicitly or implicitly) ex post information to test ex ante predictions. Adjusting predictions after making observations invalidates the test.

Returning to Equation (3), if we could observe x with no error, our total expected squared prediction error would be $E[(\tilde{y} - \beta_0 - \beta_1 x)^2] = \sigma_{\tilde{e}}^2$. When we observe x with error $\tilde{w}$, then Equation (4) provides our expected squared prediction error.

$$E[(\tilde{y} - y)^2] = E[(\beta_1\tilde{w} + \tilde{e})^2] = \beta_1^2\sigma_{\tilde{w}}^2 + 2\beta_1\sigma_{\tilde{e}\tilde{w}} + \sigma_{\tilde{e}}^2 \tag{4}$$

As is apparent from Equation (4), the observational error in x would favor our theory when $\sigma_{\tilde{e}\tilde{w}} < 0$

by partially offsetting the observational error in $\tilde{y}$. Hence, assuming errors in both x and y could make predictions look better (but not better than no error at all).

However, the extant literature often assumes[6] zero covariance between the errors, i.e., $\sigma_{\tilde{e}\tilde{w}} = 0$. In that case, of course, the expected squared prediction error increases by $\beta_1^2\sigma_{\tilde{w}}^2$ over the case when x is observed without error. Hence, errors in observing x appear to be a possible legitimate excuse for inconsistencies between theoretical predictions and observation.

This idea is wrong. Suppose there were no errors in observing x, and $\tilde{x}$ is the true x. Suppose all error is a consequence of a wrong theory (euphemistically called misspecification). Then, we are fallaciously underestimating the true error by $\beta_1^2\sigma_{\tilde{w}}^2$. The real error is all structural error. Only claiming a structural error of $\sigma_{\tilde{e}}^2$ provides false support for our theory by wrongly attributing $\beta_1^2\sigma_{\tilde{w}}^2$ to unobserved error in x.

This is not an argument against assuming error in x, which is often very appropriate for fitting models and estimating known relationships. The argument is that assuming an error in x tends to favor confirmation of our theory. It is also an argument that we should never require a researcher to assume an error in x when the purpose of the research is testing a theory, rather than estimating parameters in known relationships. Correcting for $\tilde{w}$ is a good idea when the theory is correct and a bad idea when the theory is wrong.

Giving $\tilde{w}$ a name (e.g., heterogeneity, measurement error) or providing justifications for $\tilde{w}$ only creates the illusion of fit and an excuse for errors in prediction. Any attempt to partition error into good and bad components is an illusion. Moreover, as long as $\tilde{w}$ is random and unobserved, it is unpredictable and useless for the task of validating a theoretical prediction. It would be wrong to require researchers to assume $\sigma_{\tilde{w}}^2 > 0$ when testing theories. As Friedman (1953) argues, we should evaluate models based on their predictive validity and not their assumptions.

Incidentally, the data might suggest possible modifications to the theory or new theories. Selective ex post adjustments usually favor confirmation of our ex ante theory. However, the scientific method condemns that practice, given the myriad alternative theories that perfectly predict ex post. True testing requires precise predictions before we make new observations for testing purposes.

Does β_1 Test Our Theory?

This section argues that the magnitude of estimated parameters or standardized estimates is a much weaker test of a theory (and easier to manipulate) than direct measures of the consistency of the theory with observation.

The scientific method focuses on the consistency of theory with observation. Hence, the cleanest test of a theory is the magnitude of the total error $\sigma_{\tilde{\varepsilon}}^2 = \beta_1^2\sigma_{\tilde{w}}^2 + \sigma_{\tilde{e}}^2$. However, when β_1 is unknown and estimated from the data, some researchers prefer to test theories based on the size of $|\beta_1|$ or the standardized $|\beta_1|$, sometimes because we choose the value of β_1 that maximizes the fit of the theoretical prediction. Researchers often favor this test because we choose the most favorable $|\beta_1|$ for confirming our theory. We also often use $|\beta_1|$ when we believe we must include auxiliary independent variables (e.g., instrumental variables, covariates, etc.) to adjust for still other defects in the data (euphemistically called controlling for other factors). Those variables make the model's consistency with the data a less appropriate test of the theory because predictive ability now depends on ex post independent variables (i.e., only added after observing the data and often only weakly related to the theory at risk).

Although this procedure is completely legitimate and often admirable for estimating parameters in known relationships, we should recognize that it favors confirmation of our theory. Regardless of the justification, we are still implicitly attributing prediction error to faulty data rather than to faulty theory and adjusting the data for purported defects. It is particularly dangerous when independent variables are added ex post (after observing the data) that are correlated with both x and y. In that case, these new variables can artificially improve the predicted relationship between x and y.

When we focus on $|\beta_1|$ rather than $\sigma_{\tilde{\varepsilon}}^2 = \beta_1^2\sigma_{\tilde{w}}^2 + \sigma_{\tilde{e}}^2$ there is a danger that assuming error in x will provide false support for our theory. For example, suppose we estimate β_1 by minimizing the total square error to obtain $\hat{\beta}_1$. The $\hat{\beta}_1$ estimator is inconsistent when x contains error, because its expected value underestimates β_1. To be precise, Equation (5) provides the expected value of that estimator.

$$
\begin{aligned}
E[\hat{\beta}_1] &= \beta_1\frac{\operatorname{cov}[\tilde{x},\tilde{y}]}{\operatorname{var}[\tilde{x}]} = \beta_1\frac{\operatorname{cov}[x+\tilde{w},\beta_0+\beta_1 x+\beta_1\tilde{w}+\tilde{e}]}{\operatorname{var}[x+\tilde{w}]}\\
&= \beta_1\frac{\operatorname{cov}[x+\tilde{w},\,\beta_0+\beta_1 x+\beta_1\tilde{w}+\tilde{\varepsilon}-\beta_1\tilde{w}]}{\operatorname{var}[x+\tilde{w}]}\\
&= \beta_1\left(1-\frac{\sigma_{\tilde{w}}^2}{\sigma_x^2+\sigma_{\tilde{w}}^2}\right) \qquad (5)
\end{aligned}
$$

Many researchers view Equation (5) as problematic because $|\hat{\beta}_1|$ asymptotically underestimates $|\beta_1|$. This

[6] Greg M. Allenby (personal communication) notes that methods employing instrumental variables fail to make this assumption. One could also view inclusion of instrumental variables as modifying the predicted quantity (i.e., y minus the instrumental variable). Whether that is appropriate seems to depend on the purpose of the analysis—estimation or testing a theoretical prediction.

inconsistency usually implies a bias in small samples and invalidates t-tests. Hence, several approaches such as total least squares (e.g., Nievergelt 1994, van Huffel and Vanderwalle 1991) and Bayesian procedures (e.g., Zellner 1971) have emerged for implicitly considering σ_w^2 and implicitly inflating $|\hat{\beta}_1|$. Conceptually, adjustment of $|\hat{\beta}_1|$ is possible by explicitly or implicitly assuming $\sigma_{\tilde{w}}^2 > 0$. Equation (6) shows the implicit relationship between the old $|\hat{\beta}_1|$ and the new $|\hat{\beta}_{Adjusted}|$.

$$\hat{\beta}_{Adjusted} = \hat{\beta}_1\left(1 + \frac{\sigma_{\tilde{w}}^2}{\sigma_x^2}\right) \tag{6}$$

Conceptually, as suspected error $\sigma_{\tilde{w}}^2$ increases, we inflate $|\hat{\beta}_1|$. However, we should be wary of any estimator that inflates $|\hat{\beta}_1|$ *when* we use $|\hat{\beta}_1|$ to test a theory. The reason is that we might be finding false support for our theory. For example, accounting for error (e.g., resulting from heterogeneity) in x, we might implicitly infer that $\sigma_x^2 = \sigma_{\tilde{w}}^2 = 0.5$. In that case, $\hat{\beta}_{Adjusted} = \hat{\beta}_1(1 + 0.5/0.5) = 2\hat{\beta}_1$, and we conceptually double our support for our theory. This is a conceptual argument and not a literal one because researchers need not report $\sigma_{\tilde{w}}^2$.

Equation (6) assumes that $|\hat{\beta}_1|$ underestimates $|\hat{\beta}_1|$ because of unobserved error (e.g. caused by heterogeneity) rather than a wrong theory. A wrong theory would produce the same result. Hence, if we are testing a theory, a downward biased $\hat{\beta}_1$ is doing exactly what it should. It is casting doubt on the theory. As measurement error σ_w^2 increases, $|\hat{\beta}_1|$ decreases. If the magnitude of $|\hat{\beta}_1|$, or the standardized $|\hat{\beta}_1|$, is being used as support for our theory, then there is indeed less support for our theory. A large measurement error suggests that observations fail to verify theoretical predictions, whether it be the result of measurement error or structural error or any other error. In any case, the data lack support for the theory.

There is no intent in this argument to diminish the usefulness of techniques intended for better estimating β. These clever techniques rightly make strong assumptions or better exploit the data at hand. However, one must recognize the difference between testing a theory and estimating unknown parameters in known relationships. Researchers have different objectives. If we intend some analysis to provide evidence for controversial new theories, we should avoid making assumptions that could possibly attribute structural error to observational error. Such assumptions would overstate the evidence and be contrary to a scientific investigation. Scientific investigations of new theories should be conservative and assign residual doubt to the new theory (rather than the data). In sum, it would be wrong to require researchers to assume $\sigma_{\tilde{w}}^2 > 0$ when testing theories.

Structural Equation Models

This section argues the procedural error can help validate the extraordinarily popular structural equation models (SEM).

Probably the most cited article (with over 12,000 cites[7]) in the entire marketing literature (i.e., Fornell and Larcker 1981) advocated SEM. These models take multiple measures of x to implicitly estimate σ_w^2. That is a praiseworthy procedure because multiple measures of x add information to the analysis rather than substitute assumptions for data. One concern, however, is the ease at which procedural error can increase the estimated σ_w^2.

For example, researchers might ask different survey questions and measure σ_w^2 from the consistency of the answers. However, procedural error might artificially increase σ_w^2. For example, some questions might fail to measure x. Some questions might measure different factors beyond x. Some questions might be less reliable. These procedural errors might overstate σ_w^2 and, consequently, inflate $|\hat{\beta}_1|$, providing false support for our theory.

In most scientific investigations, procedural error increases measurement error, which works against confirmation of the theory. For SEM applications, however, these procedural errors inflate σ_w^2. As we have already seen, inflating σ_w^2 favors our theory by excusing some bad predictions as measurement error rather than structural error. In sum, it would be wrong to require researchers to assume $\sigma_{\tilde{w}}^2 > 0$ when testing theories.

Unobserved Heterogeneity

This section argues that unobserved heterogeneity is simply error. When its inclusion produces dramatically different conclusions regarding a theory's validity, we must doubt the theory, the data, or both.

We know that consumers are overtly heterogeneous (e.g., Bradlow and Rao 2000) and theories based on that heterogeneity deserve attention (e.g., Blattberg et al. 1978, Hauser and Shugan 1980). More recently, however, several approaches have assumed unobserved heterogeneity. There are many ways to interpret unobserved heterogeneity. For example, Popkowski Leszczyc and Bass (1998) provide 13 different interpretations, ranging from inertia to lifestyle differences.

Suppose that our theory implies $y = \beta_0 + \beta_1 x$ but we observe $\tilde{x} = x(1 + \tilde{w})$. This is mathematically equivalent to $y = \beta_0 + \tilde{\beta}_1 x$ where $\tilde{\beta}_1 = \beta_1(1 + \tilde{w})$. As $\sigma_{\tilde{w}}^2$ approaches 0, $\tilde{\beta}_1$ approaches β_1. The model acts as if $\tilde{\beta}_1$ is a random coefficient that could be interpreted

[7] This number is from Google Scholar (4/20/2006).

as heterogeneity across the population. Equation (7) provides our modified simple linear model.

$$\tilde{y} = \beta_0 + \tilde{\beta}_1 x + \tilde{e} \tag{7}$$

Rearranging Equation (7) yields Equation (8).

$$\tilde{y} = \beta_0 + \beta_1 x + \beta_1 \tilde{w} x + \tilde{e} \tag{8}$$

If there were no heterogeneity, our total expected prediction error would be $E[(\tilde{y} - \beta_0 - \beta_1 x)^2]$. When we assume heterogeneity, then Equation (9) provides our expected squared prediction error.

$$E[(\tilde{y} - y)^2] = E[(\beta_1 \tilde{w} x + \tilde{e})^2] = \beta_1^2 \sigma^2_{\tilde{w}x} + \sigma^2_{\tilde{e}} \tag{9}$$

Here, $\sigma^2_{\tilde{w}x}$ is the variance of $\tilde{w}x$. Again, we appear to conceptually partition the error variance into two components, $\beta_1^2 \sigma^2_{\tilde{w}x}$ and $\sigma^2_{\tilde{e}}$. However, if the theory is wrong and $\sigma^2_{\tilde{w}x}$ reflects lack of consistency with the theory, then partitioning is an illusion. The total error $\beta_1^2 \sigma^2_{\tilde{w}x} + \sigma^2_{\tilde{e}}$ reflects the inconsistency between the proposed theory and the observations.

Equation (10) provides the expected value of the ordinary least squares estimator for β_1.

$$\begin{aligned} E[\hat{\beta}_1] &= \beta_1 \frac{\mathrm{cov}[\tilde{x}, \tilde{y}]}{\mathrm{var}[\tilde{x}]} = \frac{\mathrm{cov}[(1+\tilde{w})x, \beta_0 + \tilde{\beta}_1 x + \tilde{e}]}{\mathrm{var}[(1+\tilde{w})x]} \\ &= \frac{\mathrm{cov}[(1+\tilde{w})x, \beta_0 + \beta_1(1+\tilde{w})x + \tilde{\varepsilon} - \beta_1 \tilde{w} x]}{\mathrm{var}[(1+\tilde{w})x]} \\ &= \beta_1 - \beta_1 \frac{\sigma^2_{\tilde{w}x}}{\sigma^2_{\tilde{w}x} + \sigma^2_x} \end{aligned} \tag{10}$$

When x and $\tilde{w}$ are independent, Equation (11) provides the conceptual adjustment for unobserved heterogeneity.

$$\hat{\beta}_{Adjusted} = \hat{\beta}_1 \left(1 + \frac{\sigma^2_{\tilde{w}x}}{\sigma^2_x}\right) = \hat{\beta}_1 (1 + \sigma^2_{\tilde{w}}) \tag{11}$$

Conceptually, as the estimated heterogeneity $\sigma^2_{\tilde{w}}$ increases, we inflate $|\hat{\beta}_1|$. Failure to assume heterogeneity underestimates the true $|\beta_1|$. That underestimation would be problematic if Equation (7) were correct. If the theory is wrong, however, then the lower value of $|\hat{\beta}_1|$ rightfully casts doubt on the theory. The underestimation is only a concern when we have absolute faith in the theory and presume we are estimating parameters in a known relationship.

Otherwise, the lower estimate tells us that error $\tilde{w}$, if it exists, lessens the evidence supporting our theory. In sum, it would be wrong to require researchers to assume $\sigma^2_{\tilde{w}} > 0$ when testing theories. As Friedman (1953) argues, we should evaluate models based on their predictive validity and not their assumptions.

Hazard and Logit Models

This section considers hiding errors in nonlinear models—which is more insidious than linear models. As demonstrated earlier, a wrong linear model usually results in smaller coefficients for the estimated independent variables, either $|\beta_1|$ or a standardized $|\beta_1|$. Lower estimates than predicted by the theory cast doubt on the theory. This situation also often occurs for nonlinear models, but the relationship is less clear.

Simple hazard models assume a constant, stationary, fixed nonstochastic probability of an event each period. The classical example is a job seeker who tries each period to obtain employment. The probability of a match with an employer is some constant probability. An example closer to marketing is the example of a moviegoer who has a constant probability of seeing a movie each period after its release. Here, the probability might come from the researcher's uncertainty about time constraints faced by the moviegoer. Observed data will result in an estimated probability.

However, unobserved heterogeneity could give the false impression of a change in that probability over time (e.g., Vaupel et al. 1979, Lancaster 1979, Heckman and Singer 1984, Heckman 1991, Jain and Vilcassim 1991). The classical example is job seekers (e.g., Nickell 1979) who have a constant probability of finding employment over time, although some job seekers try harder than others do. For another example, some moviegoers might have a greater propensity to see the movie, but their probability is still constant over time (i.e., stationary). Assuming heterogeneity usually makes the model more consistent with the data (e.g., Elbers and Ridder 1982, Baker and Melino 2000).

However, the theory might be wrong. There might be no heterogeneity. All job seekers could be ex ante identical. As time progresses, some seekers randomly find jobs while others fail. The unlucky seekers become severely discouraged or exhaust their limited resources. Our moviegoers could be ex ante identical. However, theatrical tickets do become less desirable as DVD releases approach. Requiring the assumption of heterogeneity might hide inconsistencies between the theory and the data. It masks the need to find new theory and the fact that probabilities change over time.

Finally, consider logit, nested logit models (Hausman and McFadden 1984), and multinomial probit models (Hausman and Wise 1978) for choice data. When the theory is inconsistent with the observed data, estimated coefficients for the independent variables from the theory should approach zero, revealing no explanatory power. Assuming unobserved heterogeneity could result in inflated parameters and additional parameters that purportedly reflect that heterogeneity. These inflated parameters probably reflect better estimates when we believe the theory (estimating a known relationship). However, suppose the theory is wrong. Suppose all buyers have the same choice-probability but their consideration sets differ (van Nierop et al. 2005) or product availability differs (Hauser 1980). In that case, requiring a researcher to assume unobserved heterogeneity allows false support for a wrong theory. In sum, it is critical to determine whether the theory is valid or whether our adjustments (adding new observed and unobserved variables) are inappropriately fixing the data (e.g., observed and unobserved variables outside the theory are providing all of the explanatory power).

Maximum Likelihood Estimation and Individual Data

Unlike earlier sections of this paper, this section considers data that can potentially isolate individual differences. Suppose we have a market with 2/3 low-income people and 1/3 high-income people. We have a theory. One implication of our theory is that all high-income people buy a particular product and income dictates buying behavior. For simplicity, consider each of the following six possible purchase events as independent random draws.

Person 1 is LOW income and does NOT buy the product.

Person 1 is LOW income and does buy the product.

Person 2 is HIGH income and does NOT buy the product.

Person 2 is HIGH income and does NOT buy the product.

Person 3 is HIGH income and does buy the product.

Person 3 is HIGH income and does buy the product.

There is no apparent support for the theory in the observed data. In fact, person 2 falsifies our theory if we allow no error. There is also no relationship between income and buying behavior.

However, sometimes error can rehabilitate a theory by including some randomness. With error, person 2 alone would be insufficient to falsify the theory if other data supported the theory. That is not the case here.

Let p denote the probability a randomly chosen low-income person buys and let q denote the probability a randomly chosen high-income person buys. Assuming only a probabilistic relationship between buying and income, the likelihood function is $(4/729)(1-p)p(1-q)^2q^2$, which yields maximum likelihood estimators (MLE) $\hat{p} = \hat{q} = 1/2$ with a likelihood of 1/11664. Hence, adding randomness fails to produce support for the theory that income matters.

Now, let us include parameters for observed heterogeneity. Let r_iq denote the buy probability of a randomly chosen high-income person i. In this example, we observe only one low-income person, so we are unable to include heterogeneity for that person. However, we could construct examples with additional data that produce the same conclusion.

Consequently, the likelihood function is $(4/729) \cdot (1-p)p(1-r_2q)^2r_3q^2$, which yields maximum likelihood estimators (MLE) $\hat{p} = 1/2$, $\hat{q} = 1$, $\hat{r}_2 = 0$, $\hat{r}_3 = 1$ with a likelihood of 1/729. We improve fit by a factor of 16, i.e., 11664/729. We now have considerable support for the theory, i.e., $\hat{q} = 1$, because the heterogeneity parameter effectively gives zero weight to the inconsistent observation. However, although $\hat{q} = 1$ appears to be strong support for our theory, we could still doubt the theory because the average buy probability across all high-income people is still 1/2.

We now include parameters for unobserved heterogeneity. To keep the system identified, we add those parameters for only low-income people. Let s denote the probability that low-income people both misrepresent their income and fail to buy. Consequently, we assume unobserved heterogeneity in actual income, given a reported high income only when there is no purchase.

We now have sufficient parameters for completely heterogeneous behavior while maintaining an identified system. The new likelihood function is

$$(4/729)((1-p)(1-s))(p(1-s))((1-q+2s))^2q^2,$$

which yields MLE $\hat{s} = 1/2$, $\hat{p} = 1/2$, $\hat{q} = 1$ with a likelihood of 1/2916. Hence, by including unobserved heterogeneity, we improve fit by a factor of 4, i.e., 11664/2916 and again increase the estimated buy probability for high-income people from $\hat{q} = 1/2$ to $\hat{q} = 1$, again providing considerable support for the original theory.

To be fair, the data are consistent with the new theory that some low-income people are misrepresenting their income. However, the data might be consistent with that for a large number of ex post theories after we have seen the data, including the theory that low-income people are less likely to respond, only some high-income people always buy and high-income people are more consistent. In the end, adding parameters for unobserved variables allows more latitude for the researcher's interpretation of the data and attribution of causality. It is critical to test ex post theories (after making observations) on new data. For example, we might attempt to measure directly (now unobserved) misrepresentation or the (now unobserved) likelihood to respond.

In sum, assuming unobserved heterogeneity (or any other ad hoc parameters) could allow us to find support for possibly wrong theories. Conceptually, the variable capturing heterogeneity implicitly gives less weight to observations that are inconsistent with the original theory. In this case, the question to ask is not whether same-income consumers differ in their propensity to buy or their honesty; the question to ask is how assuming unobserved variables influences the testing procedure and the ability to falsify the theory at risk.

Of course, if we believe the underlying theory and only seek to get better estimates, incorporating unobserved heterogeneity seems to achieve remarkable results (e.g., see Chintagunta et al. 1991, Jain and Vilcassim 1991, Bradlow 1996, Allenby and Lenk 1995, Chintagunta and Prasad 1998, Chintagunta 1998, Arora and Allenby 1999, Chang et al. 1999).

Omitted Variables

This section considers the case of an omitted unobserved variable $\tilde{u}(x)$ that could influence the apparent consistency of data with theory. Suppose that $y = \beta_0 + \beta_1 x + \tilde{u}(x)$ where $\tilde{u}(x) = \gamma_0 + \gamma_1 x + \tilde{w}$ reflects the relationships between $\tilde{u}(x)$ and x. Hence, the true model is $y = (\beta_0 + \gamma_0) + (\beta_1 + \gamma_1)x + \tilde{w}$, but we estimate $\tilde{y} = \beta_0 + \beta_1 x + \tilde{e}$.

In this case, $E[\hat{\beta}_1] = \beta_1 + \gamma_1$ and, without additional information, virtually any method for estimating β_1 will be inconsistent (except $\hat{\beta}_1 = \beta_1$ everywhere) because γ_1 is completely indeterminate. If $\gamma_1 > 0$, then we overestimate β_1 and we could falsely attribute some of the influence of $\tilde{u}(x)$ to x. If $\gamma_1 < 0$, then we underestimate β_1 and we could falsely understate some of the influence of x on y because $\tilde{u}(x)$ conceals that effect.

Hence, unlike in earlier sections, not assuming an unobserved variable could possibly lead to false support for our theory. However, very few articles in *Marketing Science* have included unobserved omitted variables and found less consistency between the data and the proposed theory. Moreover, in this case, rather than assuming that heterogeneity is clouding the relationships between x and y, we are proposing an alternative theory that involves a relationship between $\tilde{u}(x)$ and y, as well as an inherently stochastic process apparently generating $\tilde{u}(x)$. Rather than modeling unobserved heterogeneity $\tilde{u}(x)$, a better approach to testing a proposed competing theory is to find an observable proxy variable for $\tilde{u}(x)$ consistent with the alternative theory.

Unobserved Heterogeneity in Behavior Experiments

This section considers unobserved heterogeneity in behavior experiments. Researchers sometimes create individual-level theories but, when collecting data, treat one observation from multiple individuals as multiple observations from one individual. Hence, the strong implicit assumption is that all individuals exhibit the same behavior, at least with respect to the theory.

If individuals exhibit different behaviors (i.e., they are heterogeneous), at least four problems appear. First, the aggregate behavior might not match the behavior of all or any of the individuals in the population (Hutchinson et al. 2000). Second, control groups fail to control for manipulations because experimental and control groups consist of different individuals. Third, if individuals differ in their behavior, single observations for each individual might be insufficient data to test theories of individual behavior. Fourth, heterogeneity in behavior across individual behavior is sometimes a competing explanation for the findings (i.e., competing with the researcher's theory).

Hutchinson et al. (2000) propose explicitly testing the assumption that individuals exhibit the same behavior when testing "reversal" effects. Their proposal has general applicability. Sometimes, it is inappropriate to replace data with assumptions (of course, sometimes it is appropriate). It is best to test theories of individual behavior with multiple observations for each individual. If tests reveal unobserved heterogeneity and that heterogeneity is an alternative explanation for findings, we have insufficient data to test the theory. Hutchinson et al. (2000) conclude, correctly, that when heterogeneity across individuals can be an alternative explanation, tests exposing that heterogeneity cast doubt on the theory but fail to fix the theory.

Problems with Holdout Samples

This section argues that holdout samples for validation are no panacea. Although econometrics provides powerful tools for marketing analysis, validation requires extraordinary care. As Stanley (1998, p. 218) notes: "An economist who is unhappy with the empirical record of his favorite theory can conduct exhaustive specification tests and searches until by luck or error a more fitting model is found or an insignificant auxiliary hypothesis can be blamed. Such is the current econometric custom."

The discussion of holdout (or saved) samples for validation requires far more attention than is possible here. Holdout samples are extraordinarily important. They are an invaluable metric for evaluating models and identifying outliers. This is particularly true when we argue that theories should rise and fall based on their predictive ability, rather than on the magnitude of estimated parameters introduced for testing purposes. Holdout samples could provide some protection from the dangers of adjusting data for defects. However, here are some issues to consider.

First, a true test of a theory requires collection of new data (i.e., ex post theory). Moreover, the scientific method dictates that we can use a sample only once, prespecifying any statistical methods before data collection (Gillies 1971, Cochran and Cox 1957) because any modifications to the methods to match the sample invalidate theory tests. Many research studies make changes based on holdout sample performance (at least during the publication review process). Although that iterative procedure might produce better estimates, it invalidates theory testing.

Second, theories often predict joint events while the holdout sample is a metric for conditional events. Hence, our theory predicts (x, y) and not other combinations, but we test the likelihood of observing y given x, i.e., $E[y \mid x]$. When we introduce new ad hoc parameters unrelated to the theory being tested, we can invalidate the test as we shift our focus to $E[y \mid x, \sigma_{\bar{w}}^2]$. Our previous MLE example illustrated the problem of focusing on conditional predictions when the conflicting information involved the joint distribution of (x, y).

Third, holdout samples are often actually partitions of a single full sample into a first and second sample. We estimate ad hoc parameters (e.g., unobserved heterogeneity) in the first sample and use those parameters to predict in the second sample. The parameters estimated in the first sample might carry sample-specific information from the first sample to test in the second sample. This information might create the illusion of consistency between the theory and the data in the second sample. This point requires some additional explanation.

If the theory is correct, then every sample will only have sampling error and holdout samples work. However, if the theory is wrong, every observation in each sample could reveal a common inconsistency between that sample and the proposed theory. For example, if the correct theory (not our proposed theory) depends on the day of the week, samples from different days will reveal different defects. Suppose we take a sample from one day and construct a holdout sample by merely partitioning it into two samples. Both of the constructed samples now reveal the same defect (not necessarily present in other samples). If we use the first sample to estimate the model, the ad hoc parameters learn and adapt to the defect. Given the second sample also has the same defect; the model might fit the holdout sample as well—providing false support for the theory. We are de facto improperly using information unique to the holdout sample to predict the holdout sample—not a valid procedure. Of course, the estimated model would not necessarily fit a truly independent sample.

Finally, our ad hoc parameters might adapt to other inconsistencies with the theory or information related to the distribution of x in the first sample. For example, we might discover when the theory makes bad predictions (e.g., for small values of x). We might implicitly weight some errors less (e.g., adopt a nonlinear model). Given that the second sample exhibits the same defects, the model appears to fit well. With hazard models, for example, the first sample might suggest adding a parameter that lessens the impact of duration dependence, which would have otherwise been evidence against the theory in the second sample. It seems inappropriate to use any parameter estimated on the first sample (which is unique to the original full sample) to predict observations in the second sample when testing theories rather than estimating established relationships. Market shares, average preferences, and other numbers estimated on the first sample are usually unique to the full sample. It is dangerous to claim better fit in the second sample after using these numbers from the first sample. Moreover, it is difficult to determine whether good predictions are a consequence of good theory or information in ad hoc parameters.

It would be worthwhile considering whether these issues are important. A better test of a theory would be to explore other implications of the theory, perhaps with different variables or different predicted relationships. Holdout samples are extraordinarily useful, and there is no intent to argue against their use. However, holdout samples might not be a panacea when we are testing theories, in contrast to estimating parameters in known relationships. It also almost seems unscientific to embrace and celebrate the tweaked theory that is least inconsistent with the data, rather than reflecting and agonizing over any obvious inconsistencies. In sum, the overall fit remains important despite performance in holdout samples.

Summary and Conclusions

Most researchers develop theories to answer research questions. For example, we know that industry standards dramatically influence markets with network externalities (e.g., Sun et al. 2004). How can a firm influence the acceptance of a standard? Many theories suggest causal relationships between variables. For example, a theory might suggest that the presence of seasonality should cause a different organizational structure to be successful. We theorize that patent laws cause an increase in innovation. Few theories include explicit error terms at the axiomatic level. Some theories have random elements, for example, agency theory. However, axioms (e.g., risk aversion) and predictions (e.g., incentive compatible arrangements always reduce the need to monitor performance) are deterministic and falsifiable. The scientific method suggests that we should reject (or, at least, revise) theories whose predictions are inconsistent with observations. Testable predictions can be prescriptive, descriptive, or both. Parenthetically, solving many marketing problems resembles games such as charades, cross-country racing, and chess more than random games such as the toss of a coin or baccarat.

Traditionally, bad predictions cast doubt on the theory. Often, we measure the credibility of a theory by examining the difference between the theory's predictions and observed outcomes. Statistical methods usually quantify that difference and refer to it as error or structural error. Recent advances in statistics create powerful tools that help accomplish many research objectives (Efron 2005). Latent class methods (e.g., Kamakura and Russell 1989) and hierarchical Bayesian frameworks have become popular (e.g., Allenby and Rossi 1998). However, as with most advances, the unintended consequences are not necessarily good. Many apparent statistical advances have created difficulties and ambiguities in the review process related to what is a good test of a theory and what is not.

There is a fundamental difference between estimating parameters in known relationships and testing a theory. For example, if a theory predicts a relationship between x and y, we must avoid introducing an ex post variable z that changes the relationship between x and y in the estimation. We might ultimately be estimating the relationship between $y-z$ and x. A true test of the theory is whether we observed the predicted relationship without numerous adjustments that favor confirmation of our theory.

Sometimes statistical tools, despite their invaluable role in estimating parameters in known relationships, focus more on error than on theory. They sometimes implicitly partition error (i.e., inconsistency between theory and data) into different components—bad structural error (i.e., defects in the theory), forgivable measurement error (i.e., errors in measuring variables), and accounting for missing variables (e.g., unobserved heterogeneity, random utility). Bad error casts doubt on a model while good error is apparently an excuse for why observations are inconsistent with theoretical predictions. For example, we might say that the theory is correct but that measurement errors or unobserved variables prevent confirmation of the theory's predictions. Alternatively, we might say that the theory is brilliant but that unobserved variables (such as unobserved heterogeneity) create defects in the data and prevent confirmation of the theory's predictions without adjusting the data.

Measurement error and missing variables should not excuse bad theory. They are only indicators that the data are insufficient to test the theory or that the theory is faulty. When theory testing, finding that different statistical assumptions about the error leads to dramatically different conclusions, we must doubt the theory, the data, or both. A test of theoretical predictions should not be extremely sensitive to assumptions related to measurement error or adjustments made to correct defects in the data. If the theory makes correct predictions, the test should be generally robust to the assumptions about the error term. The need for sophisticated assumptions about error (or unobserved variables) could indicate that the data are insufficient to test the theory or, worse, the theory is wrong.

A more valid (i.e., scientific) approach would be to consider all error as evidence against the theory. Observations should confirm theoretical predictions without elaborate adjustments to the data that account for defects in the data. Philosophically, we would prefer to eliminate error by developing better theories, observing new variables, developing more accurate measurement, and so on, rather than by adding ex post parameters to a statistical model to make the model look good (by adding unobserved variables or using information possibly unique to the data at hand). As long as unobserved variables remain unobserved, we must consider them bad error in the sense that they indicate an inconsistency between theory and observation. Although we are sometimes stuck with the data, we should focus on which observations we need to test a theory rather than on which tweak in the theory is least inconsistent with the current data. It also almost seems unscientific to embrace and celebrate the tweaked theory that is least inconsistent with the data rather than reflecting and agonizing over any obvious inconsistencies.

Unless we clearly believe that unobserved effects are present and that the inclusion of these variables works against confirmation of the theory, we should stop rejecting articles that fail to include unobserved variables. In many (if not nearly all) cases, assuming unobserved variables biases the analysis toward a better fit and, consequently, confirmation of the theory.

Unfortunately, holdout samples are not the panacea for these concerns. The scientific method requires that theoretical predictions should take place before seeking observations for testing purposes. After adjusting theory or data to account for inconsistencies between the two, we must retest using different data. It is unclear whether partitioning a data set into two parts is, de facto, creating different data. For example, when we use unique mean values (or other summary statistics) from half the data set to improve forecasts in the other half, we might be inappropriately adjusting to the unique characteristics of this one data set. Unless we are estimating unknown parameters in known relationships, we must use care that ex post adjustments (in particular, for unobserved variables) do not cloud our judgment about the validity of the theory.

Of course, in situations with limited data, limited time, limited resources, and limited theory, we need to substitute assumptions (e.g., unobserved heterogeneity) for data (e.g., observed heterogeneity). That action is perfectly reasonable (Shugan 2002). Moreover, these assumptions are perfectly appropriate for estimating established relationships. We only need to realize that we have substituted untestable assumptions for data and that there is no "free lunch."

Acknowledgments

Greg M. Allenby, Eric T. Bradlow, John R. Hauser, and Debanjan (Deb) Mitra deserve thanks for their many insightful and well-reasoned comments. There is no implied endorsement of the ideas in this editorial by these prominent scholars.

References

Allenby, Greg M., Peter E. Rossi. 1998. Marketing models of consumer heterogeneity. *J. Econometrics* **89**(1 & 2) 57–78.

Allenby, Greg M., Peter J. Lenk. 1995. Reassessing brand loyalty, price sensitivity, and merchandising effects on consumer brand choice. *J. Bus. Econom. Statist.* **13**(3) 281–289.

Arora, Neeraj, Greg M. Allenby. 1999. Measuring the influence of individual preference structures in group decision making. *J. Marketing Res.* **36**(4) 476–487.

Baker, Michael, Angelo Melino. 2000. Duration dependence and nonparametric heterogeneity: A Monte Carlo study. *J. Econometrics* **96**(2) 357–393.

Blattberg, Robert, Thomas Buesing, Peter Peacock, Subrata Sen. 1978. Identifying the deal prone segment. *J. Marketing Res.* **15**(3) 369–377.

Blaug, Mark. 1980. *The Methodology of Economics*. Cambridge University Press, New York.

Bradlow, Eric T. 1996. Negative information and the three-parameter logistic model. *J. Educational Behavioral Statist.* **21**(2) 179–185.

Bradlow, Eric T., Vithala R. Rao. 2000. A hierarchical Bayes model for assortment choice. *J. Marketing Res.* **37**(2) 259–268.

Chang, Kwang, S. Siddarth, Charles B. Weinberg. 1999. The impact of heterogeneity in purchase timing and price responsiveness on estimates of sticker shock effects. *Marketing Sci.* **18**(2) 178–192.

Chintagunta, Pradeep K. 1998. Inertia and variety seeking in a model of brand purchase timing. *Marketing Sci.* **17**(3) 253–270.

Chintagunta, Pradeep K., Alok R. Prasad. 1998. An empirical investigation of the "dynamic McFadden" model of purchase timing and brand choice: Implications for market structure. *J. Bus. Econom. Statist.* **16**(1) 2–12.

Chintagunta, Pradeep K., Dipak C. Jain, Naufel J. Vilcassim. 1991. Investigating heterogeneity in brand preferences in logit models for panel data. *J. Marketing Res.* **28**(4) 417–428.

Cochran, William G., Gertrude M. Cox. 1957. *Experimental Designs*, 2nd ed. Wiley, New York.

Dubé, Jean-Pierre, Puneet Manchanda. 2005. Differences in dynamic brand competition across markets: An empirical analysis. *Marketing Sci.* **24**(1) 81–95.

Efron, Bradley. 2005. Bayesians, frequentists, and scientists. *J. Amer. Statist. Assoc.* **100**(469) 1–5.

Elbers, Chris, Geert Ridder. 1982. True and spurious duration dependence: The identifiability of the proportional hazard model. *Rev. Econom. Stud.* **49**(3) 403–409.

Evgeniou, Theodoros, Constantinos Boussios, Giorgos Zacharia. 2005. Generalized robust conjoint estimation. *Marketing Sci.* **24**(3) 415–429.

Fornell, Claes, David F. Larcker. 1981. Structural equation models with unobservable variables and measurement error: Algebra and statistics. *J. Marketing Res.* **18**(3) 382–388.

Friedman, Milton. 1953. The methodology of positive economics. Milton Friedman, ed. *Essays in Positive Economics*. University of Chicago Press, Chicago, IL.

Gillies, Donald A. 1971. A falsifying rule for probability statements. *British J. Philos. Sci.* **22**(3) 231–261.

Golder, Peter N., Gerard J. Tellis. 2004. Growing, growing, gone: Cascades, diffusion, and turning points in the product life cycle. *Marketing Sci.* **23**(2) 207–218.

Hauser, John R. 1980. Comments on "Econometric models for probabilistic choice among products." *J. Bus.* **53**(3, Part 2) S31–S34.

Hauser, John R., Steven M. Shugan. 1980. Intensity measures of consumer preference. *Oper. Res.* **28**(2) 278–320.

Hausman, Jerry A., David A. Wise. 1978. A conditional probit model for qualitative choice: Discrete decisions recognizing interdependence and heterogeneous preferences. *Econometrica* **46**(2) 403–426.

Hausman, Jerry, Daniel McFadden. 1984. Specification tests for the multinomial logit model. *Econometrica* **52**(5) 1219–1240.

Healy, Michael J. R. 1978. Is statistics a science? *J. Roy. Statist. Soc. Ser. A* (*General*) **141**(3) 385–393.

Heckman, James J. 1991. Identifying the hand of past: Distinguishing state dependence from heterogeneity. *Amer. Econom. Rev.* **81**(2) 75–79.

Heckman, James J., Burton Singer. 1984. The identifiability of the proportional hazard model. *Rev. Econom. Stud.* **51**(2) 231–241.

Hutchinson, J. Wesley, Wagner A. Kamakura, John G. Lynch, Jr. 2000. Unobserved heterogeneity as an alternative explanation for "reversal" effects in behavioral research. *J. Consumer Res.* **27**(3) 324–344.

Jain, Dipak C., Naufel J. Vilcassim. 1991. Investigating household purchase timing decisions: A conditional hazard function approach. *Marketing Sci.* **10**(1) 1–23.

Kalnins, Arturs. 2004. An empirical analysis of territorial encroachment with franchised and company-owned branded chains. *Marketing Sci.* **23**(4) 476–489.

Kamakura, Wagner A., Gary J. Russell. 1989. Probabilistic choice model for market segmentation and elasticity structure. *J. Marketing Res.* **26**(4) 379–390.

Keuzenkamp, Hugo A., Anton P. Barten. 1995. Rejection without falsification—On the history of testing the homogeneity condition in the theory of consumer demand. *J. Econometrics* **67**(1) 103–127.

Kleindorfer, George B., Liam O'Neill, Ram Ganeshan. 1998. Validation in simulation: Various positions in the philosophy of science. *Management Sci.* **44**(8) 1087–1099.

Kluge, Arnold G. 2001. Philosophical conjectures and their refutation. *Systematic Biol.* **50**(3) 322–330.

Lancaster, Tony. 1979. Econometric methods for the duration of unemployment. *Econometrica* **47**(4) 939–956.

Little, John D. C. 1979. Aggregate advertising models: The state of the art. *Oper. Res.* **27**(4) 629–667.

Liu, Yunchuan, Z. John Zhang. 2006. The benefits of personalized pricing in a channel. *Marketing Sci.* **25**(1) 97–105.

Mahajan, Vijay, Eitan Muller. 1986. Advertising pulsing policies for generating awareness for new products. *Marketing Sci.* **5**(2) 89–106.

Montgomery, Alan L., Eric T. Bradlow. 1999. Why analyst overconfidence about the functional form of demand models can lead to overpricing. *Marketing Sci.* **18**(4) 569–583.

Nickell, Stephen. 1979. Estimating the probability of leaving unemployment. *Econometrica* **47**(5) 1249–1266.

Nievergelt, Yves. 1994. Total least squares: State-of-the-art regression in numerical analysis. *SIAM Rev.* **36**(2) 258–264.

Pearl, Judea. 2000. *Causality: Models, Reasoning and Inference*. Cambridge University Press, Cambridge, UK.

Popkowski Leszczyc, Peter T. L., Frank M. Bass. 1998. Determining the effects of observed and unobserved heterogeneity on consumer brand choice. *Appl. Stochastic Models Data Anal.* **14**(2) 95–115.

Popper, Karl R. 1965. *Conjectures and Refutations: The Growth of Scientific Knowledge*, 2nd ed. Routledge and Kegan Paul, London, UK.

Popper, Karl R. 1968. *The Logic of Scientific Discovery*, rev. ed. Hutchinson, London, UK.

Popper, Karl R. 1972. *Objective Knowledge: An Evolutionary Approach*. Oxford University Press, Oxford, UK.

Redman, Deborah A. 1994. Karl Popper's theory of science and econometrics: The rise and decline of social engineering. *J. Econom. Issues* **28**(1) 67–99.

Shugan, Steven M. 2002. In search of data: An editorial. *Marketing Sci.* **21**(4) 369–377.

Shugan, Steven M. 2005. Editorial: Are consumers rational? Experimental evidence? *Marketing Sci.* **25**(1) 1–7.

Stanley, Tom D. 1998. Empirical economics? An econometric. *J. Econom. Issues* **32**(1) 191–218.

Sun, Baohong, Jinhong Xie, H. Henry Cao. 2004. Product strategy for innovators in markets with network effects. *Marketing Sci.* **23**(2) 243–254.

Syam, Niladri B., Ranran Ruan, James D. Hess. 2005. Customized products: A competitive analysis. *Marketing Sci.* **24**(4) 569–584.

Vakratsas, Demetrios, Fred M. Feinberg, Frank M. Bass, Gurumurthy Kalyanaram. 2004. The shape of advertising response functions revisited: A model of dynamic probabilistic thresholds. *Marketing Sci.* **23**(1) 109–119.

van Helden, Albert. 1974. The telescope in the seventeenth century. *Isis* **65**(1) 38–58.

van Huffel, Sabine, Joos Vanderwalle. 1991. *The Total Least Squares Problem: Computational Aspects and Analysis. Frontiers in Applied Mathematics*, Vol. 9. Society for Industrial and Applied Mathematics, Philadelphia, PA.

van Nierop, Erjen, Richard Paap, Bart Bronnenberg, Philip Hans Franses, Michel Wedel. 2005. Retrieving unobserved consideration sets from household panel data. Working paper, Carnegie Mellon University, Pittsburgh, PA.

Vaupel, James W., Kenneth G. Manton, Eric Stallard. 1979. The impact of heterogeneity in individual frailty on the dynamics of mortality demography. *Demography* **16**(3) 439–454.

Villas-Boas, M. J. 1993. Predicting advertising pulsing policies in an oligopoly: A model and empirical test. *Marketing Sci.* **12**(1) 88–102.

Weintraub, E. Roy. 1999. How should we write the history of twentieth-century economics? *Oxford Rev. Econom. Policy* **15**(4) 139–152.

Wisan, Winifred Lovell. 1984. Galileo and the process of scientific creation. *Isis* **75**(2) 269–286.

Wu, Dazhong, Gautam Ray, Xianjun Geng, Andrew Whinston. 2004. Implications of reduced search cost and free riding in e-commerce. *Marketing Sci.* **23**(2) 255–262.

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

Zoltners, Andris A., Prabhakant Sinha. 2005. Sales territory design: Thirty years of modeling and implementation. *Marketing Sci.* **24**(3) 313–331.