Environ Resource Econ (2019) 72:559–582
https://doi.org/10.1007/s10640-017-0215-7

Reliability and Validity in Nonmarket Valuation

Richard C. Bishop[1] · **Kevin J. Boyle**[2]

Accepted: 25 December 2017 / Published online: 24 January 2018
© Springer Science+Business Media B.V., part of Springer Nature 2018

Abstract We propose a framework for assessing the accuracy of nonmarket values. This involves adapting two widely-used concepts. *Reliability* addresses variance and *validity* addresses potential biases. These concepts are formally defined and adapted to assess the accuracy of individual nonmarket valuation studies and the potential accuracy of valuation methods. We illustrate the framework by considering, in a preliminary way, the reliability and validity of the contingent-valuation and travel-cost methods.

Keywords Nonmarket valuation · Reliability · Validity · Contingent valuation · Travel-cost method

When researchers apply nonmarket valuation methods, they strive to estimate values that are accurate. But how should they judge the accuracy of their results? Or, if reviewers are assigned to evaluate the accuracy of value estimates from a nonmarket valuation study, what criteria should they bring to bear? This paper proposes a framework for assessing the accuracy of nonmarket valuation studies and methods based on the twin concepts of reliability and validity.

Nearly everything we say on the topic of accuracy has been discussed somewhere in the nonmarket valuation literature. But the literature has examined accuracy in a piecemeal fashion and evaluations of empirical studies have not been undertaken with a holistic view of accuracy. What we describe here is a framework to support more systematic and comprehensive assessments of accuracy.

✉ Kevin J. Boyle
kjboyle@vt.edu

Richard C. Bishop
rcbishop@wisc.edu

[1] Department of Agricultural and Applied Economics, University of Wisconsin-Madison, Madison, WI, USA

[2] Virginia Tech Program in Real Estate, 430A Bishop-Favrao Hall (0715), 1345 Perry Street, Blacksburg, VA 24061, USA

Googling “reliability and validity” leads to two conclusions. First, the twin concepts have been applied in many different branches of science. Second, the different sciences have adapted these general ideas to meet their specific needs. How the concepts are applied in epidemiology, for example, may be different from how they are applied in social psychology. This paper will explain how reliability and validity can be adapted to assess the accuracy of nonmarket value estimates.

Simply stated, reliability is about variance and validity is about bias. The intuition can be seen in a simple metaphor. An archer is shooting arrows at a target. Reliability of the archer depends on whether the arrows are tightly grouped or scattered about on the target. Validity depends on whether the arrows are clustered around the bull’s eye or around some other point on the target.

A valuation “method,” as we will use the term in this paper, consists of a broadly defined set of procedures used to estimate nonmarket values. Examples include the contingent-valuation, choice-experiment, travel-cost, and hedonic-price methods.

“Procedures” are the various steps used to implement a method in an empirical study. Examples of procedures include random-utility models used to implement the travel-cost method, referendum questions used to implement the contingent-valuation method, and meta-analyses used to implement the benefits-transfer method. “Individual applications” take a method and apply a specific set of procedures to estimate one or more nonmarket values.

We begin by proposing a formal framework for adapting the concepts of reliability and validity to nonmarket valuation. Reliability and validity will be investigated at two levels, the level of individual applications and the level of valuation methods. The goal is to assess the accuracy of individual value estimates. However, there is a prior question for any such assessment: is the valuation method chosen for the individual study *capable* of producing accurate results? It would make little sense to ask whether results from an individual application are accurate if the valuation method applied is known to produce highly variable and/or greatly biased value estimates. After considering how the framework applies at both the study level and the methods level, contingent valuation will illustrate how it can be applied to a stated-preference method and travel cost will exemplify how it can be applied to a revealed-preference method.

Not surprisingly, we build on the work of others. In considering the accuracy of the contingent valuation, Mitchell and Carson (1989) apply the concepts of reliability and validity and their framework has much in common with ours. Others have also picked up this theme (Loomis 1989; Whitehead et al. 2014; and others cited below). Most of the work in the nonmarket-valuation literature involving these concepts has involved stated-preference studies, largely contingent valuation. Thus, expanding to a revealed-preference method is somewhat novel.

1 Introducing the Conceptual Framework

We begin with the concept of “true value,” as the term applies in nonmarket valuation. The monetary value someone places on something is measured as either willingness to pay (*WTP*) or willingness to accept compensation (*WTA*). Consider a program that would improve environmental quality. If there is no improvement, an individual consumer (i) will enjoy some level of utility, say u_i^0. If nothing else changes, the improvement in environmental quality would put the person on a higher level of utility, $u_i^1 > u_i^0$. The true value, WTP_i^T (in this case, compensating surplus), for this change is defined as the maximum amount of money a person could pay and be exactly *indifferent* about the environmental change. The formal definition of true *WTP* for the improvement is:

$$v_i\left(P^0, Q^0, y_i\right) = v_i\left(P^0, Q^1, y_i - WTP_i^T\right)$$

where $v_i(.)$ is the consumers indirect utility function,P is a vector of prices assumed here to be constant, Q is a vector of environmental quality attributes where at least one element changes from condition Q^0 to Q^1, and y_i is income.

Let us drop the subscript i and let WTP^T represent the maximum willingness to pay of a typical or average consumer in a pre-defined population. To economize on words, this paper assumes that WTP^T is the goal of measurement, but this is done with the understanding that in some circumstances the goal could be to measure true *WTA* and that everything said about reliability and validity would continue to apply.[1] Further, the concepts discussed herein are generally applicable to environmental improvements and decrements, and to a variety of other items and services that might be valued using nonmarket valuation methods.[2]

The basic problem for reliability and validity assessment is that WTP^T is *unobservable*. It is an indifference-producing amount of money and indifference is an unobservable state of mind. In nonmarket valuation studies, choices revealed in markets and/or responses to survey questions serve as proxies for information about people's preferences.

Measurement in the social sciences often involves empirical procedures designed to quantify theoretical constructs that cannot be observed directly. An IQ test is used to try to measure human intelligence, an abstract concept. Attitude questions in surveys try to measure phenomena such as racial prejudice and self-esteem, mental states that cannot be observed directly. Mathematics exams aim to measure competence in mathematics, an unobservable concept. Likewise, travel-cost models, contingent-valuation studies, and the other nonmarket valuation methods provide estimates of unobservable true values people hold for changes in items that are not directly valued in markets.

Since human intelligence, racial prejudice, WTP^T, and other such concepts are unobservable, researchers seek empirically tractable, but indirect, ways to assess accuracy. This is *reliability and validity assessment*.

Suppose the research problem is to evaluate *WTP* for an improvement in water quality in a river. Let WTP^{T} represent the true *WTP* of a typical household for this improvement.

The researchers collect data on property sales in the area affected by the water quality improvement and apply a hedonic property-value model to estimate the marginal value for a typical household. Let this be WTP_j^E, which represents the estimated average value per household for a small improvement in water quality. In theory, at least, the hedonic model could be estimated repeatedly with resampling and WTP_j^E represents the value estimate for the *jth* application. The relationship between the true value and the *jth* estimated value can be written as:

$$WTP_j^E + e_j = WTP^T \;\forall\; j,$$

where e_j is the error in estimating the true value in the *jth* application, which can be positive, negative, or zero. Accuracy, therefore, deals with the magnitude of e_j. Each time a new estimate of WTP_j^E is obtained through resampling, it has a corresponding value of e_j.

[1] There are some additional issues that must be addressed when estimating WTA. For a discussion, see Kim et al. (2015).

[2] Here we take an implied assumption that a public good is being valued and recognize that for private goods the incentives for truthful preference revelation are different.

Estimation usually involves errors, so it is unrealistic to expect that $e_j = 0, \forall\ j$. It is important to separate normal sampling variation from additional variation in estimates due to the valuation method and procedures used. Let us partition the error term as $e_j = e_{rj} + e_{mpj}$ where e_{rj} denotes the random error associated with sampling from a population and e_{mpj} denotes measurement error associated with the estimation method (m) and procedures (p) used.

Since WTP_j^E is an estimated mean, the standard error of this estimate is:

$$se_{WTP_j^E} = s_j \big/ \sqrt{n_j}$$

where $\mathrm{s_j}$ is the standard deviation and $\mathrm{n_j}$ is the sample size. The larger the standard error, the less reliable is the estimate.

For purposes of exposition, let us partition the standard deviation into two sources of variation:

$$s_j = s_{rj} + s_{mpj}.$$

Here, s_{rj} is the standard deviation due to sampling variation and s_{mpj} is the standard deviation associated with the valuation method chosen and the procedures selected to estimate WTP_j^E. For example, consider a contingent-valuation study where the scenario contains a confusing description of the item to be valued, which leads subjects to make their own assumptions about what is being valued, increasing the dispersion of the value estimates across the sample. This would reduce reliability, but it might or might not bias the estimated mean. Increasing the sample size would reduce $se_{WTP_j^E}$ and thus enhances reliability independent of the source of variation (random or systematic), but reliability may also be improved by avoiding procedures that increase dispersion in value estimates (s_{mpj}).

If appropriate sampling procedures are used, $E\left(e_{rj}\right) = 0$, and bias occurs if $E\left(e_{mpj}\right) \neq 0$. In the hedonic model example, bias could occur, for example, when an omitted relevant variable is correlated with an included variable used to estimate *WTP*. With contingent valuation, this might occur if an incentive compatible elicitation format was not used and respondents provided valuation responses that systematically over or under estimated WTP^T. Bias is reduced by selecting procedures that have been shown to minimize confounding effects of procedures on value estimates.

Both reliability and validity are important to accuracy. Following the archery metaphor in the introduction, real world researchers typically get only one or a few "shots" at the "value" target, i.e., most of the time, they can only apply the procedures once or a few times in each situation. Hence, the more erratic are value estimates, the less confidence researchers can have in results from any one application, even if the estimates tend, on average, to be "centered" on the true *WTP*. That is, it is possible for estimated values to be unbiased (centered on the true value), but have a large variance. Or, if estimates are biased (tend to be different from WTP^T), then value estimates can still have a small variance. The goal is to discover and apply nonmarket valuation procedures that are both reliable (minimize variance) and valid (minimize bias).

In real-world applications, some bias is typically unavoidable in nonmarket valuation research or in all empirical economic estimates for that matter. As a practical matter, then, the absence of bias is not a strict criterion for assessing validity. Likewise, reliability is not a clear cut yes or no condition. As we shall stress again in the next section, reliability and validity are normally matters of degree in terms of how much bias and variability are acceptable or unacceptable. This depends of the policy issue or other context where the values will be used.

We now have the bare bones of the framework we propose for assessing the accuracy of nonmarket value estimates. The next section fleshes out the framework, a necessary process before we illustrate how to apply it using the contingent-valuation and travel-cost methods as examples. Since WTP^T is unobservable, we explore how indirect evidence is used to assess the accuracy of nonmarket values.

2 Reliability and Validity Assessment

As noted above, reliability and validity occur at two levels, the level of individual applications and the level of methods. Much of what we know about the accuracy of methods comes from individual applications. Here we discuss these concepts at the study level and then generalize to methods.

2.1 Reliability Assessment for Individual Applications

Reliability assessment for an individual application involves more than whether the econometric estimator chosen minimizes variance; it includes all procedures in a study that can influence the magnitude of variance even when a variance-minimizing estimator is applied.

It follows that the first step in reliability assessment is to consider the estimated standard errors of the value estimates. The larger the estimated standard error of a value estimate, the less reliable it is deemed to be.

In the broader social science literature, further insights into reliability are typically gained from test–retest experiments where a set of procedures for a method is applied at time t and then replicated at time $t + 1$ (Carmines and Zeller 1979; Zeller and Carmines 1980). A test–retest experiment is like taking two shots at the target. While two shots are not many shots at the target, if the two estimates are statistically indistinguishable, the measure is deemed reliable.[3] In nonmarket valuation studies involving surveys—e.g., stated-preference studies and most travel-cost studies—the survey could be administered at two points in time. For methods that involve building data sets from sources other than surveys—most notably, many hedonic studies—data could be gathered at two points in time. Such reliability investigations would be predicated on the assumptions that preferences, market conditions, and other factors that might influence value estimates have not changed between t and $t + 1$.

Investigating reliability in nonmarket valuation studies can be tricky. For studies involving surveys, for example, a test–retest design could have the same people retake the survey at $t + 1$. For this to work, the observations at the two points in time should be independent, but this may be difficult to accomplish. If the time between the surveys is short, people may mimic their t responses at $t + 1$. As the time between treatments increases, there may be fundamental changes that could affect preferences or income. For example, using contingent valuation, Loureiro and Loomis (2017) found a significantly higher value for the effects of an oil spill in Spain when the survey was done in 2006 compared to when it was redone in 2009. However, they attribute the difference to changes in macroeconomic circumstances (by 2009, Spain was deep into a recession), and not to unreliability.

Note that studies with small and large variances can be judged reliable using a test–retest investigation if the null hypothesis of no difference in value estimates between time t and $t+1$

[3] By consequence, this also provides evidence that the chosen valuation method and implementation procedures in a nonmarket valuation study are reliable, but reliability of a method, as discussed below, requires multiple reliability studies before general conclusions can be reached about the method and implementation procedures. Similar considerations apply to validity assessment.

cannot be rejected. More broadly, a large variance may make it more difficult to reject test–retest reliability even if the two measures are quite different, and a small variance may make it more difficult to confirm reliability when the two measures are similar. Thus, test–retest experiments are evidence of reliability or unreliability, but are not a litmus test.

In closing this section, the social science literature sometimes views reliability and test–retest reliability as synonyms. Because test–retest is simply an assessment procedure and there are challenges such as those outlined above, it would be unfortunate if this usage carried over to nonmarket valuation research. Test–retest is simply a procedure for investigating reliability.

2.2 Validity Assessment for Individual Studies

Given that true values are unobservable, indirect evidence of validity (i.e., evidence that does not require that true values be known) must be used. Three types of validity are commonly considered in social science research: content, construct, and criterion validity (Carmines and Zeller 1979; Zeller and Carmines 1980). We refer to these as the *Three Cs* of validity assessment.

As applied here, *content validity* focuses on whether the valuation method chosen, and all procedures used to implement it are conducive to measuring the true value. This concept of content validity is broader than elsewhere in the nonmarket valuation literature. For example, Mitchell and Carson (1989) focus on the adequacy of the contingent-valuation survey instrument. We are suggesting a definition that goes beyond to include all procedures from the initial definition of value through the reporting of value estimates. Content validity assessments are based on economic and econometric theory, accumulated experience of the analyst and others working in the field, and evidence from past studies.

In choice experiments, for example, Holmes et al. (2017) have suggested that there are seven implementation steps: (1) characterize the decision problem; (2) identify and describe the attributes to be valued; (3) develop the experimental design; (4) develop the survey; (5) implement the survey; (6) estimate the model; and (7) interpret the results for policy analysis or other applications. Content-validity assessment of a study using a choice experiment would ask whether the procedures used at each of these steps were fully conducive to estimating the true value.

Moving from content to construct validity, the focus shifts to the study's estimation results. *Construct validity* begins with prior expectations about how the true value ought to be related to other variables. Such prior expectations are motivated by theory, intuition, and past empirical evidence. They are translated into hypotheses that can be tested using the study's data and statistical results. For example, holding other things constant, one would expect the price of a dwelling unit to be related to its square footage. Should a hedonic price function fail to show such a relationship, this is grounds for questioning the study's construct validity.

A common type of construct-validity test is convergent validity (e.g., Hanley 1989). Convergent validity considers whether two or more valuation methods that are designed to estimate the same true value provide statistically similar value estimates.[4] If values from the two methods are statistically indistinguishable, this is evidence of construct validity. At the same time, it is not possible to conclusively say the value estimates are valid, because

[4] Another form of construct validity testing used to evaluate the performance of econometric models is cross validation. Here, a hold-out sample of observations is used to evaluate model performance. The estimated model is used to predict responses of the hold-out sample. Predicted responses are graphed as a function of observed responses, and valid responses should fall on or near a 45-degree line from the origin. Here we do not discuss all approaches used to investigate construct validity, but use convergent validity as an example because it is the most commonly used assessment procedure in the literature.

both value estimates could be biased in the same direction. Likewise, finding a statistically significant difference in value estimates is not conclusive evidence of invalidity. The degree to which either study is invalid is not known because the true value is not known. It may be that one valuation study produced a biased estimate of value and the other did not, or that both studies provided biased estimates of value, e.g., one might have provided an overestimate and the other might have produced an underestimate. Thus, if the estimates are statistically indistinguishable, this is suggestive evidence, but not confirmation, that both methods have produced unbiased estimates.

Since there are limitations on the inferences that can be drawn from construct-validity tests, some researchers have turned to tests of criterion validity. *Criterion validity* involves tests comparing results from two valuation methods, one method with uncertain validity and another method that is widely accepted as having a high level of validity. Results from the widely-accepted method serve as the "criterion" (the proxy for the true value) to evaluate the accuracy of results from the method with uncertain validity.

Consider an example. List and Gallet (2001) summarize the literature on what has come to be known as "hypothetical bias" in stated-preference studies. In laboratory or field settings, results from stated-preference studies have been compared with results from transactions involving the actual exchange of money for goods or services. Since most economists would be satisfied if values from stated-preference studies were comparable to values from well-functioning markets, results from the actual transactions serve as the criterion for judging the validity of results from the parallel stated-preference applications. If stated-preference studies perform well in comparison with cash transactions, this is evidence of criterion validity.

Confusion occasionally arises about whether such investigations are tests of criterion-validity or tests of convergent validity. For example, if we called comparisons of contingent-valuation and travel-cost estimates convergent-validity tests, we would be implicitly assuming that neither is considered more valid than the other. In contrast, in the criterion-validity example, the outcomes from cash transactions have presumed validity while stated-preference outcomes have uncertain validity.

Some have rightly asked, why are the outcomes cash experiments deemed to be "more" valid as they may have confounding experimental design and implementation effects? If the presumed validity does not hold, then the comparison is simply an investigation of convergent validity. Thus, while neither construct nor criterion-validity tests are perfect, criterion-validity tests are taken as stronger indicators of true values because the criterion measures are presumed to be unbiased or, at a minimum, contain much less bias.

Of course, if the method with widely accepted validity could be broadly applied at reasonable cost, then it would be used and there would be no need for the method with uncertain validity. However, the more accepted method may be too expensive or may be usable only under special circumstances. Thus, studies using the criterion method are normally not amenable to many empirical applications.

As can be seen from the brief overviews of the *Three Cs*, each provides a different form of insight regarding validity and each has limitations. This is the reason we propose a weight-of-evidence approach for interpreting the results of validity investigations in the next section.

2.3 Reliability, the Three Cs, Judgment, and the Weight of Evidence

Without being able to observe true values, there are no "litmus" criteria to say how much variability makes a study unreliable and how much bias makes a study invalid. That is, variance and bias are statistical concepts, whereas reliability and validity are judgments imposed on statistical observations. Three points follow. First, reliability and validity are matters of degree. Estimated values are not reliable or unreliable and valid or invalid per se. Rather, some values are more reliable and/or more valid than others. Second, disagreements about reliability and validity often arise because different individuals apply different standards about what is and is not acceptable. Third, what is consider reliable and valid in one context may not be in another context.

Sound judgments about reliability and validity involve careful attention to the weight of all the evidence. The smaller the estimated standard error of the value estimate, the more confident one can be that the value estimate is reliable. The strongest validity inferences will apply to studies that were implemented using procedures with established content validity, that have passed appropriate construct-validity tests, and, most importantly, that have employed a method and related procedures that have proven strong in criterion-validity testing. The *Three Cs* can be thought of as three legs of a stool. Judgments about the validity of value estimates should rest on all three legs. If one or more legs are weak, then validity suffers. For example, a study with strong content validity may be judged less valid if prior expectations are not met in convergent or other construct-true values tests.

We would emphasize at this point that biased estimates are useful. This is particularly true when the bias is small and the direction of bias can be inferred. An obvious example from benefit-cost analysis would be where an estimate of individual benefit is thought to be an underestimate, yet still aggregate benefits exceed costs. When valuation studies are conducted to support decision making, the debate about accuracy should center on how much bias and unreliability are tolerable before the value estimates cease to be useful from a practical decision-making standpoint in the context of specific applications.

2.4 The Reliability and Validity of Methods

One way to proceed is to think about how the framework could be applied during the evolution of a new method. As the first few attempts to apply the new method become available, preliminary evidence regarding reliability could be gleaned by examining estimated standard errors. Likewise, early construct- and/or criterion-validity testing of new methods would yield preliminary insights about whether procedures have been successful or not. These early insights, in turn, would be implemented and tested in more studies as the basis for content validity. As the process continues, the three legs that a successful method rests upon would become progressively stronger—or the method would be abandoned. In this way, judgments about the reliability and validity of a method evolve and become stronger, or not, as evidence accumulates across many studies.

Hence, the reliability of a method should be judged based on whether past studies, employing state-of-the-art procedures, have produced value estimates with estimated standard errors that are within acceptable limits. Test–retest experiments provide evidence to support such insights.

In the same way, a method can be judged as more or less content valid based on whether a group of researchers in the field agree on a set of procedures that can be used to estimate true values to a satisfactory approximation. Consensus among researchers is based to a large degree on repeated testing documented by construct- and criterion-validity studies published in the peer-reviewed literature. As the literature accumulates, controversies are inevitable because there are few hard guidelines on how preference data should behave, and this is an area where different individuals have imposed differing criteria regarding what is and is not valid.

A method should be judged to be construct valid based on the extent to which an accumulated body of results from individual applications is consistent with prior expectations based on theory and intuition, and the consistency of findings across construct- and criterion-validity studies. Some failures of construct-validity tests in individual applications are not damning. Such failures may stem from peculiarities in the circumstance of the errant study, statistical flukes, going beyond the domain where the method can be applied, or other causes. On the other hand, if the method in question fails such tests frequently, then, at a minimum, this would point toward the conclusion that current, state-of-the art "procedures" are not construct valid. Perhaps selected procedures need to be evaluated and revised. If the method fails in many procedural dimensions of individual applications, it would be deemed to lack construct validity and require great improvement or be abandoned. What one is looking for in construct-validity assessment of methods is the weight of evidence about what works well and where *inherent* flaws or limitations arise that most researchers in the field would agree should not be present. Insight on the directional effect of potential biases is important evidence for such weight of evidence evaluations.

If criterion-validity tests are feasible, they can provide potent evidence of the validity of a nonmarket valuation method. When a study employing state-of-the-art procedures yields value estimates that compare favorably to estimates from a widely-accepted criterion, researchers feel much greater confidence in the method in question. Poor performance in criterion validity tests can be damning. But, there are some caveats to this seemingly simple formula for validity assessment.

First, criterion-validity failure in a few studies may not be fatal. There may be a danger that the criterion studies are flawed, which could lead to a false negative conclusion. Or, the results may be context dependent. That a valuation method performs poorly in one context should not be taken as evidence that it will fail in other contexts. Criterion-validity studies must be replicated to see if positive or negative results are robust.

Secondly, a criterion-validity investigation is a summative investigation of a specific outcome, say estimated WTP for a nonmarket good or service, but these tests may say little about how specific steps in implementing a study enhance or diminish validity. It is possible that one step leads to over estimation and another to under estimation, with offsetting effects. Such internal anomalies within the valuation process would not necessarily be detected in a criterion validity investigation. Thus, criterion-validity testing is not a litmus test and requires accompanying construct-validity tests investigating specific procedures.

Thirdly, to say that a body of researchers achieves a level of consensus about the reliability and validity of a method does not mean that there will be complete agreement. A method can only continue to evolve and improve if those developing the method debate how to apply it correctly and set priorities for further research.

2.5 Other Concepts of Validity

Before proceeding, we should point out that others in the social sciences sometimes address validity in terms different from the *Three Cs*. For example, Trochim (2002) builds a framework around the concepts of conclusion validity, internal validity, construct validity (defined differently than here), and external validity. Such concepts have carried over into experimental economics. For example, Roe and Just (2009, pp. 1266–1267) speak of "internal validity" ("the ability of a researcher to argue that observed correlations are causal"), external validity ("the ability to generalize the relationships found in a study to other persons, times, and settings"), and ecological validity ("the extent that the context in which subjects cast their decisions is similar to the context of interest"). More specific to the stated-preference literature, Diamond (1996) and Diamond and Hausman (1994) consider "internal consistency tests," such as the adding-up test, as an approach to investigating validity of the contingent-valuation method. Vossler and coauthors, in a series of publications, often use the term "external validity" (e.g., Vossler et al. 2003); Vossler and Kerkvliet 2003; Vossler et al. 2012).

The alternative binning of such validity concepts overlaps with and differs from the *Three Cs*. For example, internal validity assessment, including internal consistency tests overlaps with construct-validity assessments as we use that term. Yet, not all construct-validity investigations address internal validity as defined by Roe and Just; some construct-validity test outcomes are simply observed relationships without causal relationships known or inferred, e.g., convergent-validity investigations.

External validity deals with whether value estimates can be generalized. For example, when researchers attempt to judge whether nonmarket value estimates for a sample can be generalized to a population, they are considering the external validity of their results. In such a case, external validity may be enhanced by a probability sampling plan (content validity) and by testing survey results on demographics against demographics for the population (construct validity). Further, the ability to generalize outcomes of *Three Cs* investigations to other study applications, what we refer as the weight of evidence, might also be interpreted as a type of external validity. In contrast, Vossler and coauthors appear to use the external validity and criterion validity as synonyms.

In sum, internal validity and external validity focus on some of the same issues as the *Three Cs*, and there is overlap between criterion validity and ecological validity as well. Our goal here is not to resolve all the different concepts and conceptual groupings of validity. Rather we attempt to bring order to the standard approaches to validity in the nonmarket valuation literature, the *Three Cs*. Nonetheless, considering the same issues from different vantage points can be useful, as we shall see in when the focus turns to the contingent-valuation method.

From the foundation we have outlined, the rest of this chapter is devoted to considering how reliability and validity might be applied in evaluating two well-known valuation methods, contingent valuation and the travel-cost method. We ask whether studies applying the contingent-valuation and travel-cost method using state-of-the-art procedures yield reliable and valid estimates of *WTP*.

We hope that readers will consider what follows a "test drive" of the framework, not our attempt at final assessments of the methods under scrutiny. If some readers disagree with what we say about one or both methods, we will have achieved our goal. Scientific judgments will differ, and our purpose is not to have the last word, but to illustrate a more comprehensive framework for conducting such debates and to motivate more refined reliability and validity research, and more informed judgments on reliability and validity across all nonmarket-valuation methods.

3 Accuracy of the Contingent Valuation Method

We are using contingent valuation (CV) as a case example for stated-preference methods, but the accuracy framework has general applicability to other stated-preference methods. Further, the goal here is not to provide a comprehensive assessment with all possible citations, but to provide a general flavor of the literature on reliability and validity of CV using supporting examples from the literature.

The CV method has been quite controversial since its inception, but the debate has been more focused for the past 25 years. In preparing for litigation in the aftermath of the *Exxon*

Valdez oil spill in 1989, both the federal government and the State of Alaska (Carson et al. 2003) commissioned CV studies to estimate damages to the public. After settlement of the case, Exxon's team published a number of critiques of CV aimed at questioning the accuracy of the method (Hausman 1993; Diamond and Hausman 1994). The controversy has continued into the 21st Century with one journal article's title succinctly expressing the views of the critics: "Contingent Valuation: From Dubious to Hopeless" (Hausman 2012). Such a sweeping conclusion, based on limited and selective evidence, does not adequately represent the weight of evidence bearing on the reliability and validity of the CV method. Nor do the many methodological CV studies published individually provide a comprehensive assessment. Thus, a comprehensive and systematic assessment, based on reliability and the *Three Cs,* is needed to understand the accuracy of CV.

3.1 Contingent Valuation Reliability

Standard errors of CV estimates of value are often calculated, and we know of no cases where they have been considered problematical. Furthermore, many studies have performed test–retest experiments and found evidence for concluding that well-conducted CV studies produce reliable estimates of value (Berrens 2000; Carson et al. 1997; Jones-Lee et al. 1985; Kealy et al. 1988, 1990; Loehman and De 1982; Loomis 1989, 1990; Reiling et al. 1990; Brower and Bateman 2005). Early in the process, investigators began to recognize that resurveying the same subjects might be misleading if subjects remember their first responses when completing the second survey. However, Teisl et al. (1995) find that when an independent control group was used, test–retest reliability continued to hold. Further, McConnell et al. (1998) address this question econometrically and conclude that, in their case, correlations between earlier and later responses is due to heterogeneous preferences, not recalling earlier responses. We know of no studies that contradict these findings.

3.2 Contingent Valuation Content Validity

Moving to a discussion of content validity, we begin by asking whether a body of researchers (economists and other social scientists) has achieved consensus around CV procedures that will lead to satisfactory approximations of true values. The large literature on CV (Carson 2011) demonstrates such a consensus. Researchers in the field would agree on the basic procedures involved in doing a state-of-the art study, including the general efficacy of carefully defining the changes to be valued, deciding whose values are to be counted, defining a suitable sample size, and so on (Johnston et al. 2017). And there seems to be substantial agreement on many of the details under each step. For example, few would question the desirability of subjecting survey drafts to qualitative research and pretesting, framing CV questions as single-bounded dichotomous choices posed as referenda where possible, and collaborating with scientists outside of economics (Arrow et al. 1993; Mitchell and Carson 1989). A wide consensus has also developed regarding appropriate econometric approaches (e.g., Haab and McConnell 2002).

At the same time, there are choices that investigators must make routinely where there is no clear-cut guidance from theory or the literature and these choices can influence value estimates. Three examples come to mind.

First, the issue of which payment vehicle to use persists. Many recent studies have used a tax increase as a payment vehicle. Given the aversion many people have to taxes, many have argued that using a tax vehicle yields value estimates that tend to be conservatively low compared to true values, all else equal. A tax may be the most realistic and consequential payment vehicle for public provision of many items. This illustrates how a biased estimate may still be useful. Nevertheless, the use of tax vehicles remains an uneasy compromise (Morrison et al. 2000). Content validity suffers when bias is introduced, whether doing so is intentional or unintentional. And there are situations where taxes are implausible or otherwise unworkable. What then?

Second, the appropriate frequency of payments and their duration is not at all clear (Stevens et al. 1997). Many have explored this issue in focus groups and found that it matters to people whether payments are monthly or annual and whether they extend over one year or some other period. This is not simply a matter of discounting at some conventional rate (Stumborg et al. 2001).

The third example is the issue of whether to estimate individual or household values (Lindhjem and Navrud 2009). While conventional theory focuses on the individual "consumer," it is obvious that many households engage in joint budgeting and decision making.

There is currently no clear guidance from theory or from empirical studies about how to resolve such design dilemmas. The content validity of CV applications is enhanced somewhat by clearly documenting investigator decisions with support from qualitative research (e.g., focus groups) and robustness analyses of response data, where possible, to understand the potential magnitude of the impacts on value estimates.

3.3 Contingent Valuation Construct Validity

Economic theory in its most basic form suggests hypotheses that can be tested when the CV method is applied. The first might be termed "negative price sensitivity." Stated differently, for a given change in quantity or quality the percentage of people who would pay for the change should not increase as the cost increases? Few if any CV studies have failed this price sensitivity test.

The second test suggested by basic theory might be termed "positive income sensitivity." Many CV studies have found that income has a positive effect on values. For example, Jacobsen and Hanley (2009), in a meta-analysis of *WTP* for biodiversity, found that values rise with income and that the income elasticity of *WTP* is less than one. We would also expect income elasticity to vary across environmental resources, which Jacobsen and Hanley found.

Scope tests are another theory-based, construct-validity test.[5] Surveys of studies that have conducted scope tests (Carson et al. 1997; Desvousges et al. 2012) show that enough studies do pass to show that scope test failure is not endemic to the CV method. Hence, when individual studies fail income sensitivity and/or scope tests, this may raise questions about the construct validity of the individual studies, but not about the CV method itself.

In other construct-validity tests, many researchers have found significant, intuitively plausible relationships between contingent values and respondent environmentalism, use of the resource, and other such variables. Such variables can often be interpreted as proxies either for preferences or for differences in individual circumstances. Widespread demonstrations of expected relationships between *WTP* and these proxy variables have encouraged the conclusion that CV studies tap into economic preferences, which supports the construct validity of the method (Carson et al. 2001).

A number of studies have tested for the convergent validity of CV and revealed-preference values. Carson et al. (1996) conducted a meta-analysis of studies that contained both CV and revealed-preference estimates of value based on travel-cost, hedonic, and averting-cost methods. They conclude that CV values tend to be somewhat less than revealed-preference

[5] Another alternative is the "adding up test", which can be viewed as a special case of a scope test.

values. For example, under one set of assumptions, the average ratio of CV to revealed-preference values is 0.89, with a 95% confidence interval of [0.81–0.96] and a median of 0.75. Other studies comparing CV with revealed-preference values including Shrestha and Loomis (2001, 2003), Gen (2004), and Brander et al. (2007), also found that CV estimates tend to be smaller.

To us, this is an instance where the proverbial glass may be half (or more) full, rather than half empty. Yes, it would be cleaner if the two methods consistently yielded the same values. But, given that the two approaches are so different in procedural dimension, the fact that value estimates come out close seems to us to be good news about the validity of CV. This conclusion is further reinforced by the Carson et al. (1996) finding that the revealed-preference measures and corresponding CV measures are strongly correlated. This suggests that revealed-preference and CV methods are measuring the same underlying values, though some bias may still be present.

3.4 Contingent Valuation Criterion Validity

Two types of criterion-validity tests have been conducted, those using simulated markets and those using actual voting behavior. In simulated market experiments, values based on actual cash transactions serve as the validity criterion for comparison with results from CV. In the voting comparisons, CV studies using a referendum format are conducted and the results compared to actual referendum outcomes as the validity criterion. Both approaches can provide important insights.

Three meta-analyses, using much of the same data, but different analysis procedures, compare CV results to results from parallel simulated markets (List and Gallet 2001; Little and Berrens 2004; Murphy et al. 2005). All three conclude that CV estimates tend to exceed simulated market estimates. For example, Murphy et al. use a data set from 28 studies and compare the ratio of the CV values to corresponding simulated market values. The ratios range from 0.76 to 25.08, with a mean of 2.60, a standard deviation of 3.52, and a median of 1.35.[6] While details vary, the other meta-analyses showed roughly the same results. The conclusion is that the CV method is subject to a positive hypothetical bias.

But how strong is the case for hypothetical bias, really? The CV treatments in the hypothetical bias experiments very often stress that the valuation exercises were hypothetical. Stressing that nothing will happen makes the CV treatments in these experiments inconsequential (Carson and Groves 2007). Stating that subjects will not, under any circumstances, actually have to pay renders them incentive incompatible (Carson et al. 2014). Further, many of these studies were conducted with private goods where the incentive properties are not known. All three of these characteristics of the hypothetical bias studies contrast markedly with typical state-of-the-art CV applications. In actual studies, consequentiality is supported by stressing that policy makers will consider results in deciding what to do about the issues raised in the CV scenario. Incentive compatibility is supported by using incentive compatible mechanisms and payment vehicles that could be implemented and require subjects to pay if the change being valued is provided to consumers. They also focus on public or quasi-public goods under field conditions.

[6] These numbers should be interpreted with caution. One suspects that several of the studies included in the meta-analyses would not fare well in a comprehensive validity assessment. Still the consistency of overvaluation by CV cannot be ignored.

The meta-analyses of hypothetical bias studies have not addressed these aspects. A growing body of evidence supports the hypothesis that hypothetical bias does not occur in well-designed CV exercises that are incentive compatible, consequential, and have binding payments (Carson et al. 2014; Poe and Vossler 2011; Vossler et al. 2012). More research building on these insights should be a high priority.

The hypothetical bias literature highlights another issue that is commonly encountered in the social sciences, namely the potential conflict between internal and external validity.[7] Laboratory experiments can have substantial internal validity because experimenters have control over the structure of the experiments. In the hypothetical bias studies, researchers can control the setting of the experiment, the instructions that subjects receive, the incentives they face, and other parameters. This should enhance their ability to approximate true values within the experimental context. On the other hand, achieving internal validity can make laboratory experiments contrived and artificial compared to the real world, raising doubts about external validity. To be sure, failure of contingent values to measure up in the lab creates doubts about their validity in the real world. However, results from the lab that satisfy internal validity do not necessarily generalize to the real world external validity of CV applications.

In a "natural" experimental setting, results from CV exercises formatted as referendum questions are compared with the outcomes of actual referenda. This provides an advantage over lab experiments in testing the criterion validity as an artificial market is not needed for the criterion counterfactual. We know of five voting comparisons studies that support the validity of the CV method (Mitchell and Carson 1989; Champ and Brown 1997; Vossler and Kerkvliet 2003; Vossler et al. 2003; Johnston 2006). In all these studies, CV performed well compared to actual voting. This appears to be a promising area for further investigation.

3.5 The Weight of Evidence on Contingent Valuation Accuracy

More than two decades ago, the U.S. Blue Ribbon Panel on Contingent Valuation was asked to assess the accuracy of CV results for purposes of assessing monetary damages from spills of oil and various other toxics into the environment. It concluded,

> Contingent valuation studies convey useful information. We think it is fair to describe such information as reliable[8] by the standards that seem to be implicit in similar contexts, like market analysis of new and innovative products and the assessment of other damages normally allowed in court proceedings. Thus, the Panel concludes that contingent valuation studies can produce estimates reliable enough to be the starting point of a judicial process of damage assessment, including lost passive-use [non-use] values (Arrow et al. 1993, 4610).

Our conclusion from reviewing the updated, contemporary evidence is consistent with that of the NOAA Panel. The CV method stands up rather well to the scrutiny fostered by the framework proposed here. We believe that the case is strong enough to justify the use of CV to support decision making, not only in litigation, but in policy analysis more generally.

Regarding reliability:

- Reliability has been demonstrated in multiple studies, including test–retest experiments.

The CV method has also stood up well in validity tests.

[7] We are grateful to an anonymous reviewer for pointing this out.

[8] The use of the term "reliable" by the panel follows the U.S. legal system's use of this term and can be considered a synonym to our use of the term "accuracy," which includes reliability, as defined here, and validity.

- Researchers working in the area have reached substantial agreement on procedures to evaluate content validity.
- Construct-validity tests support the validity of the method.
- Criterion-validity tests have proven supportive so far.

Despite this evidence, there are professionals in several disciplines who continue to be suspicious of the CV method because it uses stated-preference data. However, a holistic view of the weight of evidence in reliability and validity testing should allay many of their fears. In fact, where limitations have been identified the research has often demonstrated the direction of the effects (e.g., potential bias), a result that can be used to support decision making. This is not to say that the CV method is a finished product. As we have stressed, more research is needed across a broad front. We have highlighted needs relating to some specific design procedures, and consequentiality, incentive compatibility and criterion validity.

Furthermore, concluding that the CV method has established enough reliability and validity to be useful does not mean that any one application can be automatically considered accurate. Rather, the focus shifts to whether the application in question can demonstrate reliability and content and construct validity. This is in keeping with the NOAA Panel, which made itemized recommendations that they considered essential for accuracy of individual CV studies. For example, CV studies should format their valuation questions as referenda and include reminders of subjects' budget constraints. Nearly all NOAA Panel recommendations, as updated by Johnston et al. (2017), fit nicely with the framework espoused in this paper and are incorporated in the best studies today. In sum, while it should no longer be necessary for an individual study to defend the CV method as a general approach, researchers continue to have the burden of addressing key issues of reliability and validity that are likely to be most relevant to their specific applications. This can be done through pretesting, data validity and robustness investigations, and reliance on insights from the literature.

4 Accuracy of the Travel-Cost Method

Using our framework, we now consider whether the travel-cost (TC) method, when applied using state-of the-art procedures, can produce reliable and valid estimates of true values. As with the CV example, the goal here is not to provide a comprehensive assessment and all possible citations, but to provide a general flavor of the literature. We are using the TC method as a case example of revealed-preference methods and the framework can be applied to hedonic and averting-behavior methods.

TC practitioners have been concerned with accuracy issues going back to some of the earliest studies (Brown and Nawas 1973; Cesario and Knetsch 1970; Stevens 1969; Trice and Wood 1958). However, to our knowledge, past research has not provided a systematic, comprehensive reliability and validity assessment of the TC method. In fact, we would argue that this shortcoming applies to all revealed-preference methods.

Relative to CV, TC has been relatively free of controversy. We postulate that this is due to the prima facie acceptance of revealed-preference data, with perhaps a blind eye to the many necessary investigator choices involved in any TC application.

As with CV, what we say should be considered a first attempt to evaluate what the literature implies regarding the reliability and validity of the TC method. Others will no doubt have much more to say.

4.1 Travel Cost Reliability

The lack of unacceptably large standard errors in value estimates from most TC studies supports reliability.

Some additional insights can be gleaned from a handful of studies on recall of recreation behavior. Several studies have reported recreational participation rates from independent samples varying in the length of recall time. The U.S. Fish and Wildlife Service conducts a periodic National Survey of Fishing, Hunting, and Wildlife-Associated Recreation, which traditionally asked subjects to recall their recreational participation over the past year. Chu et al. (1992) found that this annual recall led to increased variance of reported recreation participation relative to semiannual and quarterly recall periods. Gems et al. (1982) found that longer recall periods resulted in this same pattern of results for marine sport anglers who took four or more trips per year when 2-week and 2-month recall periods were compared, but the variance did not increase for non-avid anglers (three or fewer trips). In another study of recreational fishing the standard deviation of days fished more than doubled for 3- and 6-month recall periods versus immediate recall (Tarrant et al. 1993). In these types of recall studies, it is presumed that data reported for shorter recall periods will have the same or less variability. The collective findings from these studies suggest that using recreation participation data with long periods of recall may tend to increase the variance of reported participation, which may reduce the reliability of TC estimates of value, all else being equal. We will return to the topic of "recall bias" in the subsection on content validity.

Mazurkiewicz et al. (1996) conducted a test–retest experiment using immediate and 4-month recall for a 1-week moose hunt, and found that the 4-month recall period did not affect the variance of hunter participation. This finding does not necessarily contradict the finding of Chu et al., Gems et al., and Tarrant et al. because the moose hunt is a 1-week hunt. This is not like fishing, where there can be many trips over several months.

Still, recall studies do not answer the fundamental reliability question of whether TC studies conducted over time with multiple samples from the same population would produce approximately the same value estimates. Test–retest studies by Bhattacharjee et al. (2009) and Parsons and Stefanova (2011) support reliability, but Mkwara et al. (2015) question this conclusion using a test–retest study with more advanced econometric analysis of the data.

4.2 Travel Cost Content Validity

Is there broad agreement among researchers on the procedures to be used in TC studies to achieve at least minimal accuracy? Evidently, the answer is yes. TC is far from being considered an experimental method that has yet to prove itself (Herriges and Kling 2008; Ward and Beal 2000). Rather it stands as a high-ranking tool among revealed-preference methods. This seems to support the conclusion that the TC method, when applied using state-of-the-art procedures, has high content validity.

Still, there are several of what Parsons (2017) has termed "soft spots" in current practice. Parsons suggests that more realistic ways of valuing travel time are needed. This issue is made thornier by the possibility that recreationists may be heterogeneous in how they treat time.[9] He also points out that overnight trip and multiple-purpose trip modeling needs to be improved. In addition, practical, realistic models of intertemporal substitution need to be developed to account for the possibility that, if one decides not to visit a selected site

[9] This possibility was pointed out to us by an anonymous reviewer. This is an interesting point that we do not recall seeing it in the literature.

today, one may, instead of going to a substitute site, visit the preferred site later. Furthermore, standard practices have not yet been developed to measure out-of-pocket trip costs.

To Parson's list we would add other loose ends related to endogeneity, allocation of joint costs, and recall bias. TC models tend to assume that trip costs (distance from home, cost per mile, travel time, vehicle and other durable equipment costs, and lodging and subsistence costs) are treated as exogenous by subjects, whereas they may involve choices that are correlated with recreational behavior. As a case in point, consider the choice of where one lives, a determinant of trip expenses and time spent in travel. At its heart, the TC method uses the behavior of those with higher travel costs to predict participation rates of people with lower travel costs if the "price" of visits were raised. What if people with different travel costs also have different preferences regarding the recreational activity in question? If some people choose where they live based in part on nearness to recreation sites, a serious measurement error could be introduced. Recreation activities like rock climbing, fly fishing, and downhill skiing come to mind as examples. Other things being equal, the result would be to underestimate values (see Parsons 2017), but other things may not be equal. For example, those living closer to the site may pay a premium for housing if the site is close to environmental amenities.[10] In addition, how to allocate joint costs of durables like vehicles and equipment to individual trips continues to be intractable.

In addition, the research cited in the subsection on TC reliability raises the possibility of over reporting if data with longer recall periods are used. Chu et al. (1992) found that annual recall led to over-reporting of activity relative to semiannual and quarterly recall periods for recreational fishing. Gems et al. (1982) found that longer recall periods resulted in this same pattern of results for marine sport anglers who took four or more trips per year when comparing 2-week and 2-month recall periods. Mazurkiewicz et al. (1996), however, found that a 4-month recall period did not affect mean hunter participation for a 1-week moose hunt. As was the case in the TC reliability discussion above, the specific activity and limited participation time of the moose hunt does not contradict the findings from fishing studies. Over reporting of participation can lead to an upward bias in value estimates

So, while up to a point, a case can be made for the content validity of the TC method, much work remains to be done to firm up the soft spots. In the extreme, Randall's conclusion two decades ago may still apply (Randall 1994, p. 88)

> … traditional TCM yields only ordinally measurable welfare estimates. The household production function formulation of TCM 'resolves' this problem only by imposing severe and un-testable analytical restrictions. TCM cannot serve as a stand-alone technique for estimating recreation benefits; rather, it must be calibrated using information generated with fundamentally different methods.

These soft spots mean that the content validity of the TC method, as currently practiced, is not as strong as some have supposed. The soft spots must be acknowledged by carefully reporting of assumptions used and robustness tests of critical assumptions, but these affirmative actions are not a panacea for more developmental research.

4.3 Travel Cost Construct Validity

As with CV, the construct validity of the TC method can be considered from two different angles. First, can it demonstrate expected relationships between value estimates and priors

[10] Other variables, such as fish catch rates reported in surveys, may also be endogenous.

drawn from theory and intuition? And second, do TC values demonstrate convergence with values derived from other methods?

As for prior expectations, TC applications consistently find a negative relationship between travel costs and participation, satisfying negative price sensitivity (downward sloping demand functions).

Income sensitivity is more interesting. Do values from TC models show a consistent tendency toward increasing value estimates with increases in subjects' incomes, all else equal? This question is not often asked. We have not done a complete survey of the literature, but within a sample of articles in the peer reviewed literature since 2000, most studies did not include income as a possible explanatory variable (Boxall and Adamowicz 2002; Haener et al. 2004; Hynes et al. 2007; Kinnel et al. 2006; Landry and Hindsley 2011; Lupi et al. 2003; Moeltner and Englin 2004; Scarpa et al. 2008; Timmins and Murdock 2007). Of the exceptions, some found positive, significant income sensitivity (Boxall et al. 2003; Landry and Liu 2009); others found no significant effect of income (Massey et al. 2006); and some had more than one model with mixed results (Grijalva et al. 2002; Murdock 2006). While it would be surprising if a thorough investigation failed to demonstrate widespread positive income sensitivity, for now, construct validity has not been confirmed in this dimension.

The counterpart of CV scope tests is the many TC studies that have examined how values change with site quality. These studies often found statistically significant relationships that imply differences in values (Bockstael and Kling 1988; Bockstael et al. 1987; Egan et al. 2009; Hanley et al. 2003; Kaoru 1995; Massey et al. 2006; McConnell 1977; Parsons and Stefanova 2011; Phaneuf 2002; Joen et al. 2011) This is evidence for the construct validity of the TC method.

Meta-analyses provide additional support by documenting that value estimates demonstrate expected relationships to many independent variables. For example, Smith and Kaoru (1990) found that TC value estimates vary systematically with the type of recreation activity valued (see also Shrestha and Loomis 2003). This evidence also supports the construct validity of TC value estimates.

Regarding convergent validity, recall the discussion regarding the comparisons of CV and revealed-preference methods, including the TC method. Carson et al. (1996) find that TC estimates are 10% larger than CV estimates on average. While we concluded that convergence has not been confirmed, TC and CV comparison studies have produced value estimates that are sufficiently close to provide mutually reinforcing evidence of convergent validity.

4.4 Travel Cost Criterion Validity

There is not much evidence of TC method criterion validity, pro or con. We know of only one study comparing TC values with simulated market values. McCollum (1986) compares TC estimates for a deer hunting experience with WTP estimates from simulated markets for the same deer hunting opportunity. The main finding was that behavior captured in TC models was not statistically distinguishable from the behavior in cash markets when relatively low values for travel time (10% to 33% of the wage rate) were used.

More research is needed, and, in the meantime, lack of criterion-validity studies weakens this leg of the three-legged stool to support the validity of the TC method.

4.5 The Weight of Evidence on Travel-Cost Accuracy

What can be said by way of conclusions about the reliability and validity of the TC method?

Regarding reliability:

- Lack of unacceptable standard errors across many studies supports the reliability of TC estimates of WTP.
- Reliability will likely be enhanced by short recall periods when collecting data.
- Research is needed using test–retest methods.

There is substantial support for the validity of the TC method:

- It is rooted in revealed-preference data.
- There is broad agreement on procedures.
- It is supported in several ways by results from construct-validity testing.

At the same time, research is needed to address some gaps.

- "Soft spots" force researchers to make assumption that lack theoretical and empirical support.
- Income sensitivity has not been adequately explored.
- Criterion validity has not been established.

We speculate that the confidence economists have in revealed preference data may have lulled TC researchers into complacency regarding the validity of assumptions that must be applied to collect and analyze TC data and the effects of these assumptions on value estimates. All said, in the absence of the intense scrutiny that has been applied to CV, the TC literature has accumulated an impressive record of use in support of decision making. However, we challenge researchers to conduct more, innovative research to confirm and enhance accuracy.

5 Conclusions

The accuracy framework we describe can serve as an outline to assist investigators in systematically enhancing the reliability and validity of nonmarket valuation studies and users of value estimates to assess the accuracy of studies. If the framework is viewed as a skeleton, research to date has done much to flesh out the body, but the work is not finished.

As we have seen, the debate over the accuracy of nonmarket valuation methods has been most intense when focused on the CV method. Much has been accomplished since Scott (1965, p. 37) cynically asserted, "ask a hypothetical question and you get a hypothetical answer."[11] We have argued that the weight of evidence suggests that CV has sufficient reliability and validity to be a useful tool to inform policy analysis and litigation, although research should continue to improve the method. The rigorous scrutiny that has focused on CV needs to continue and expand to the other nonmarket valuation methods.

While the TC method has not been subjected to the aggressive scrutiny of the CV method, there is sufficient reliability and validity research to support the use of this valuation method as well. All the same, there is a clear need for reliability research and more validity research to address the soft spots in the method.

Reliability and validity of nonmarket valuation studies could be enhanced if more attention were given to five issues. First, we have assessed the reliability and validity of the CV and TC methods in a preliminary way, more work is needed to flesh out these assessments and to apply the framework to other nonmarket-valuation methods.

Second, thinking about the evolution of nonmarket valuation, we would speculate that a disproportionate share of the research has focused on improving econometric estimation

[11] Interestingly, if by "hypothetical," Scott meant "inconsequential," he has a point.

methods compared to other issues relating to study design and execution. Advanced econometric methods cannot address fundamental issues in the design of studies where information is missing to identify the effects. We believe that wide application of the accuracy framework introduced here would help to identify and balance research priorities. Given that advanced econometric methods are applied to address data limitation, improvements in the design and execution of nonmarket valuation studies can relax restrictive conditions imposed on econometric estimators.

Third, criterion-validity investigations can provide the most potent tests of nonmarket valuation methods. Our assessment of the TC method illustrates the need to extend criterion-validity research to other nonmarket valuation methods. More of this research will provide clear insights to advance the methods.

Fourth, we suggest that a weakness of economics as a discipline compared to other disciplines is that, once a study is published, replications are very hard to publish. In other disciplines, publication of replication studies is a normal part of scientific progress. The consequence of this disciplinary parochialism is that our base of knowledge may be broad, but not very deep. This undermines the ability to assess the accuracy of economic research methods. While we believe that this applies throughout economics, it seems particularly true of nonmarket valuation.

Lastly, progress in improving the reliability and validity of nonmarket valuation methods has been slowed by lack of funding for methodological research, both from scientific funding agencies and from the agencies that use nonmarket values. Thus, too much of research on methods depends on investigators' ability to cobble together funding to conduct small experiments or to attach their methodological research to practical, policy-oriented studies for which funding has been more generally available. This results in too few studies to address fundamental issues of study design and implementation.

Despite these issues, the CV and TC examples illustrate the remarkable advances that have been made in stated- and revealed-preference methods of nonmarket valuation over the last 50 some years. For CV, a recent example is the use of incentive compatible question formats in the context of consequential valuation exercises. For TC, this might be the use of RUM models to more explicitly account for substitutes and effectively include resource quality in estimated models. Thus, returning to the glass of water metaphor, we postulate that overall the nonmarket valuation glass is at least half full and gaining volume. Our accuracy framework can enhance these gains when applied in a systematic and balanced fashion to guide future research.

To the extent that others disagree with us on specific points, this is healthy for the evolution of nonmarket valuation. In terms of assessing accuracy, disagreements can enhance understanding and point to needed research so long as we all strive to go into the debate with an open mind and a commitment to consider the weight of all the evidence.

In closing, we reemphasize that some bias is inherent in nearly all empirical estimates regardless of the discipline. When weighing the evidence for specific applications, researchers should consider how much variance and bias are acceptable to support the relevant decision-making context. We hope that our message of unfinished business (the glasses are not full) will stimulate more reliability and validity research on all nonmarket-valuation methods to enhance the information provided to support decision making.

Acknowledgements This article used Chapter 12 in Champ et al (2017) as a starting point. We thank Springer, the book's publisher, for permission to do so. We also thank Ian Bateman for encouraging us to adapt and further develop the ideas in the chapter. Two anonymous reviewers provided many helpful comments and suggestions. Coral Bishop provided invaluable help in preparing the manuscript.

References

Arrow K, Solow R, Portnoy PR, Leamer EE, Radner R, Schuman H (1993) Report of the NOAA panel on contingent valuation. Fed Reg 58:4601–4614

Bhattacharjee S, Kling CL, Herriges JA (2009). Kuhn-Tucker estimation of recreation demand—a study of temporal stability. Agricultural and Applied Economics Association annual meeting

Berrens RP (2000) Reluctant respondents and contingent valuation surveys. Appl Econ Lett 7:263–266

Bockstael NE, Kling CL (1988) Valuing environmental quality: weak complementarity with sets of goods. Am J Agric Econ 70(3):654–662

Bockstael NE, Hanemann WM, Kling CL (1987) Modelling recreational demand in a multiple site framework. Water Resour Res 23(5):951–960

Boxall PC, Adamowicz WL (2002) Understanding heterogeneous preferences in random utility models: a latent class approach. Environ Resour Econ 23(4):421–446

Boxall PC, Rollins K, Englin J (2003) Heterogeneous preferences for congestion during a wilderness experience. Resour Energy Econ 25:177–195

Brander LM, Van Beukering P, Cesar HSJ (2007) The recreational value of coral reefs: a meta-analysis. Ecol Econ 63(1):209–218

Brower R, Bateman IJ (2005) Temporal stability and transferability of willingness to pay for flood control and wetland conservation. Water Resour Res 46(2):353–361

Brown WG, Nawas F (1973) Impact of aggregation on the estimation of outdoor recreation demand functions. Am J Agric Econ 55(2):246–249

Carmines EG, Zeller RA (1979) Reliability and validity assessment. In: Sage university paper series on quantitative applications in the social sciences. Sage, Woburn, MA

Carson R (2011) Contingent valuation: a comprehensive bibliography and history. Edward Elgar, Cheltenham

Carson R, Flores NE, Martin K, Wright JL (1996) Contingent valuation and revealed preference methodologies: comparing the estimates for quasi-public goods. Land Econ 72:80–99

Carson R, Flores NE, Meade NF (2001) Contingent valuation: controversies and evidence. Environ Resour Econ 19:173–210

Carson R, Groves RM (2007) Incentive and informational properties of preference questions. Environ Resour Econ 37:181–210

Carson R, Groves T, List JA (2014) Consequentiality: a theoretical and experimental exploration of a single binary choice. J Assoc Environ Resour Econ 1:171–207

Carson R, Hanemann M, Kopp RJ, Krosnick JA, Mitchell RC, Presser S, Rudd PA, Smith VK, Conaway M, Martin K (1997) Temporal reliability of estimates from contingent valuation. Land Econ 73(2):151–163

Carson R, Mitchell RC, Hanemann M, Kopp RJ, Presser S, Ruud PA (2003) Contingent valuation and lost passive use: damages from the Exxon Valdez oil spill. Environ Resour Econ 1:171–207

Cesario FJ, Knetsch JL (1970) Time bias in recreation benefit estimates. Water Resour Res 6(3):700–704

Champ PA, Brown TC (1997) A comparison of contingent and actual voting behavior. In: Proceedings from W-133 benefits and cost transfer in natural resource planning, 10th Interim Report, pp 77–98

Champ PA, Boyle KJ, Brown TC (eds) (2017) A primer on nonmarket valuation, Springer, New York

Chu A, Eisenhower D, Hay M, Morganstein D, Neter J, Waksberg J (1992) Measuring the recall error in self-reported fishing and hunting activities. J Off Stat 8(1):19–39

Desvousges WH, Matthews K, Train K (2012) Adequate responsiveness to scope in contingent valuation. Ecol Econ 84:121–128

Diamond P (1996) Testing the internal consistency of contingent valuation surveys. J Environ Econ Manag 30(3):337–347

Diamond PA, Hausman JA (1994) Contingent valuation: Is some number better than no number? J Econ Perspect 8:45–64

Egan KJ, Herriges JA, Kling CL, Downing JA (2009) Valuing water quality as a function of water quality measures. Am J Agric Econ 91(1):106–123

Gems B, Ghosh D, Hitlin R (1982) A recall experiment: impact of time on recall on recreational fishing trips. In: Proceedings of the section on survey research methods. American Statistical Association, Washington, DC

Gen S (2004) Meta-analysis of environmental valuation studies. Public Policy, Georgia Institute of Technology, Atlanta

Grijalva TR, Berrens RP, Bohara A, Shaw W (2002) Testing the validity of contingent behavior trip responses. Am J Agric Econ 84(2):401–414

Haab TC, McConnell KE (2002) Valuing environmental and natural resources: the econometrics of non-market valuation. Edward Elgar, Cheltenham

Haener MK, Boxall PC, Adamowicz WL, Kuhnke DH (2004) Aggregation bias in recreation site choice models: resolving the resolution problem. Land Econ 80(4):561–574

Hanley N (1989) Valuing non-market goods using contingent valuation. J Econ Surv 3(3):235–252

Hanley N, Schlapfer F, Spurgeon J (2003) Aggregating the benefits of environmental improvements: distance-decay functions for use and non-use values. J Environ Manag 68:297–304

Hausman JA (1993) Contingent valuation: a critical assessment. Emerald Group Publishing Limited, Bingley

Hausman JA (2012) Contingent valuation: from dubious to hopeless. J Econ Perspect 26:43–56

Herriges JA, Kling CL (2008) Revealed preference approaches to environmental valuation, volumes I and II. Routledge, New York

Holmes T, Adamowicz W, Carlsson F (2017) Choice experiments. In: Champ PA, Boyle KJ, Brown TC (eds) A primer on nonmarket valuation, Ch. 5. Springer, New York

Hynes S, Hanley N, Garvey E (2007) Up the proverbial creek without a paddle: accounting for variable participant skill levels in recreational demand modelling. Environ Resour Econ 36(4):413–426

Jacobsen JB, Hanley N (2009) Are there income effects on global willingness to pay for biodiversity conservation? Env Resour Econ 43(2):137–160

Joen Y, Herriges JA, Kling CL, Downing JA (2011) The role of water quality perceptions in modelling lake recreation demand. In: Bennett JW (ed) The international handbook on non-market environmental valuation. Edward Elgar, Cheltenham, pp 74–101

Johnston RJ (2006) Is hypothetical bias universal? Validating contingent valuation response using a binding public referendum. J Environ Econ Manag 52:469–481

Johnston R, Boyle KJ, Adamowicz W, Bennett J, Brouwer R, Cameron TA, Hanemann WM, Hanley N, Ryan M, Scarpa R, Tourangeau R, Vossler CA (2017) Contemporary guidance for stated preference studies. J Assoc Environ Resour Econ 4(2):319–405

Jones-Lee MW, Hammerton M, Philips PR (1985) The value of safety: results of a national sample survey. Econ J 95:49–72

Kaoru Y (1995) Measuring marine recreation benefits of water quality improvements by the nested random utility model. Resour Energy Econ 17(2):119–136

Kealy MJ, Dovidio JF, Rockel ML (1988) Accuracy in valuation is a matter of degree. Land Econ 64(2):158–171

Kealy MJ, Montgomery M, Dovidio JF (1990) Reliability and predictive validity of contingent valuation: Does the nature of the good matter? J Environ Econ Manag 19:244–263

Kim Y, Kling CL, Zhao J (2015) Understanding behavioral explanations for the WTP-WTA divergence through a neoclassical lens: implications for environmental policy. Ann Rev Resour Econ 7(1):169–187

Kinnel JC, Bingham MF, Mohamed AF, Desvousges WH, Kiler TB, Hastings EK, Kuhns KT (2006) Estimating site choice decisions for urban recreators. Land Econ 82(2):257–272

Landry CE, Hindsley P (2011) Valuing beach quality with hedonic property models. Land Econ 87(1):92–108

Landry CE, Liu H (2009) A semi-parametric estimator for revealed and stated preference data—an application to recreational beach visitation. J Environ Econ Manag 57(2):205–218

Lindhjem H, Navrud S (2009) Asking for individual or household willingness to pay for environmental goods? Environ Resour Econ 43(1):11–29

List JA, Gallet CA (2001) What experimental protocol influence disparities between actual and hypothetical stated values? Evidence from a meta-analysis. Environ Resour Econ 20:241–254

Little J, Berrens RP (2004) Explaining disparities between actual and hypothetical stated values: further investigation using meta-analysis. Econ Bull 3:1–13

List JA, Gallet CA (2001) What experimental protocol influence disparities between actual and hypothetical stated values? Environ Resour Econ 20(3):241–254

Loehman E, De VH (1982) Application of stochastic choice modeling to policy analysis of public goods: a case study of air quality improvements. Rev Econ Stat 54:474–480

Loomis JB (1989) Test–retest reliability of the contingent valuation method: a comparison of general population and visitor responses. Am J Agric Econ 71:76–84

Loomis JB (1990) Comparative reliability of the dichotomous choice and open-ended contingent valuation techniques. J Environ Econ Manag 18(1):78–85

Loureiro ML, Loomis JB (2017) How sensitive are environmental valuations to economic downturns? Ecol Econ 140:235–240

Lupi F, Hoehn JP, Christie GC (2003) Using an economic model of recreational fishing to evaluate the benefits of sea lamprey (*Petromyzon marinus*) control on the St Marys River. J Great Lakes Res 29:742–754

Massey DM, Newbold SC, Gentner B (2006) Valuing water quality changes using a bioeconomic model of a coastal recreational fishery. J Environ Econ Manag 52(1):482–500

Mazurkiewicz SM, Boyle KJ, Teisl MF, Morris KI, Clark AG (1996) Recall bias and reliability of survey data: moose hunting in Maine. Wildl Soc Bull 24(1):140–148

McCollum DW (1986) The travel cost method: time, specification and validity. University of Wisconsin-Madison, Madison

McConnell KE (1977) Congestion and willingness to pay: a study of beach use. Land Econ 53(2):185–195

McConnell KE, Strand IE, Valdes S (1998) Testing temporal reliability and carry-over effect: the role of correlated responses in test–retest reliability studies. Environ Resour Econ 12:357–374

Mitchell RC, Carson RT (1989) Using surveys to value public goods: the contingent valuation method. Resources for the Future, Washington, DC

Mkwara L, Marsh D, Scarpa R (2015). Testing the stability of welfare estimates in travel cost random utility models of recreation: an application to the Rotorua Lakes, New Zealand. 59th Australian Agricultural and Resource Economics Society Conference

Moeltner K, Englin J (2004) Choice behavior under time-variant quality: state dependence versus Play-It-By-Ear in selecting ski resorts. J Bus Econ Stat 22(2):214–224

Morrison MD, Blamey RK, Bennett JW (2000) Minimising payment vehicle bias in contingent valuation studies. Environ Resour Econ 16:407–422

Murdock J (2006) Handling unobserved site characteristics in random utility models of recreation demand. J Environ Econ Manag 51(1):1–25

Murphy JJ, Allen PG, Stevens TH, Weatherhead D (2005) A meta-analysis of hypothetial bias in stated preference valuation. Environ Resour Econ 30:313–325

Parsons GR (2017) Travel cost. In: Chap PA, Boyle KJ, Brown TC (eds) A primer on nonmarket valuation, Ch. 6, Springer, New York

Parsons GR, Stefanova S (2011) Gauging the value of the short-term site closures in a travel-cost RUM model of recreation demand with a little help from stated preference data. In: Whitehead J, Haab TC, Huang J (eds) Preference data for environmental valuation: combining revealed and stated preference approaches. Routledge, Taylor & Francis Group, New York

Phaneuf DJ (2002) A random utility model for total maximum daily loads: estimating the benefits of watershed-based ambient water quality improvements. Water Resourc Res 38(11):1254

Poe GL, Vossler CA (2011) Consequentiality and contingent values: an emerging paradigm. Edward Elgar, Cheltenham

Randall A (1994) A difficulty with the travel cost method. Land Econ 70(1):88–96

Reiling SD, Boyle KJ, Phillips ML, Anderson MW (1990) Temporal reliability of contingent values. Land Econ 66(2):128–134

Roe BE, Just DR (2009) Internal and external validity in economics research: tradeoffs between experiments, field experiments, natural experiments, and field data. Am J Agric Econ 91(5):1266–1271

Scarpa R, Thiene M, Train K (2008) Utility in willingness to pay space: a tool to address confounding random scale effects in destination choice to the Alps. Am J Agric Econ 90(4):994–1010

Scott A (1965) The valuation of game resources: some theoretical aspects. In: Fisheries Canadian, Report, iv. Department of Fisheries of Canada, Ottawa, Canada

Shrestha RK, Loomis JB (2001) Testing a meta-analysis model for benefit transfer in international outdoor recreation. Ecol Econ 39(1):67–83

Shrestha RK, Loomis JB (2003) Meta-analytic benefit transfer of outdoor recreation economic values: testing Out-of-Sample convergent validity. Environ Resour Econ 25(1):79–100

Smith VK, Kaoru Y (1990) Signals or noise? Explaining the variation in recreation benefit estimates. Am J Agric Econ 72(2):419–433

Stevens JB (1969) Effects of nonprice variables upon participation in water-oriented outdoor recreation: comment. Am J Agric Econ 51(1):192–193

Stevens TH, DeCoteau NE, Willis CE (1997) Sensitivity of contingent valuation to alternative payment schedules. Land Econ 73(1):140–148

Stumborg BE, Baerenklau KA, Bishop RC (2001) Nonpoint source pollution and present values: a contingent valuation study of Lake Mendota. Appl Econ Perspect Policy 23(1):120–132

Tarrant MA, Manfredo MJ, Bayley PB, Hess R (1993) Effects of recall bias and nonresponse bias on self-report estimates of angling participation. North Am J Fish Manag 13(2):217–222

Teisl MF, Boyle KJ, McCollum DW, Reiling SD (1995) Test-retest reliability of contingent valuation with independent sample pretest and posttest control groups. Am J Agric Econ 77:613–619

Timmins C, Murdock J (2007) A revealed preference approach to the measurement of congestion in travel cost models. J Environ Econ Manag 53(2):230–249

Trochim, William M (2002). The research methods knowledge base. http://www.anatomyfacts.com/research/researchmethodsknowledgebase.pdf. Accessed 30 July 2017

Trice AH, Wood SE (1958) Measurement of recreation benefits. Land Econ 34(3):195–207

Vossler CA, Doyon M, Rondeau D (2012) Truth in consequentiality: theory and field evidence on discrete choice experiments. Am Econ J Microecon 4:145–171

Vossler CA, Kerkvliet J (2003) A criterion validity test of the contingent valuation method: comparing hypothetical and actual voting behavior for a public referendum. J Environ Econ Manag 45:631–649

Vossler CA, Kerkvliet J, Polasky S, Gainutdinova O (2003) Externally validating contingent valuation: an open space survey and referendum in Corvallis, Oregon. J Econ Behav Org 51:261–277

Ward FA, Beal D (2000) Valuing nature with travel cost models. Edward Elgar, Cheltenham

Whitehead J, Noonan DS, Marquardt E (2014) Criterion and predictive validity of revealed and stated preference data: the case of Mountain Home Music concert demand. Econ Bus Lett 3(2):87–95

Zeller RA, Carmines EG (1980) Measurement in the social sciences: the link between theory and data. Cambridge University Press, Cambridge