Open Access

Advances in Environmental and Engineering Research

LIDSEN Publishing Inc.

Review

The Reliability of Willingness-to-Pay Responses in the Contingent Valuation Method: A New Single-Sample Approach

Bradley Jorgensen *

La Trobe University, Melbourne, Australia; E-Mail: b.jorgensen@latrobe.edu.au

* **Correspondence:** Bradley Jorgensen; E-Mail: b.jorgensen@latrobe.edu.au

Academic Editor: Grigorios L. Kyriakopoulos

Special Issue: Environmental Economics and Management

Adv Environ Eng Res
2023, volume 4, issue 1
doi:10.21926/aeer.2301009

Received: October 24, 2022
Accepted: January 12, 2023
Published: January 18, 2023

Abstract

The Contingent Valuation (CV) Method, like other stated preference techniques, seeks to measure economic preferences for public goods. Throughout the development and application of this non-market procedure, the accuracy of the measured preferences has been front-and-centre among practitioners and potential users. The most important issue of debate has been the extent to which the method can reliably measure economic preferences. In this article, a new methodology is described that enables multiple indicators of latent preferences. Multiple-indicator CV (MCV) enables the application of reliability analyses that are well established in psychology and sociology and represent the foundation of evaluating the measurement of latent variables. Furthermore, with the new MCV approach, the reliability of measurement at the individual level can be assessed in a single administration of the MCV survey thereby alleviating any need for longitudinal methodologies or comparison of mean estimates with other valuations of the same ecosystem service or public good should these be available. Once adequate reliability is established, the multiple-indicator framework supports the estimation of mean values via existing econometric techniques. With greater confidence in the reliability of measured contingent values, the interpretation of validity tests is enhanced.

Keywords
Stated preferences; contingent valuation; reliability; validity; willingness to pay

1. Introduction

Environmental economics has developed the contingent valuation (CV) method to establish the value of ecosystem services. The relevance of the method stems from the absence of markets in which transactions can be observed and used to infer the maximum amount that individuals are willing to pay (WTP) or willing to accept (WTA) compensation for. Transactions involving buyers and sellers represent observed behaviour from which economic preferences – a latent (or unobservable) variable [1] – can be understood. Without a means of estimating preferences, efficiency analyses such as the benefit-cost test cannot be undertaken.

Accurately measuring latent variables is not straightforward because they are unobservable constructs. The accuracy of the values of latent variables that are inferred from observed indicators depends upon the degree to which the measured data coincides with the unobservable construct. One source of influence on indicators is *random measurement error* [2]. This type of error in the observed data reflects solely chance variation. Random sources of variation serve to decrease the correlation between the latent variable and its indicators.

To index the degree of random error in observed indicators, the term *reliability* is used [3]. Reliability refers to the extent to which an indicator consistently provides the same result under the same conditions. This consistency does not mean that individuals should give the same response on one or more occasions. Rather, it requires that the rank-order of individual responses remains the same from one valuation occasion to the next. Larger amounts of random variance decrease an indicator's reliability and contributes to its inconsistent performance in situations where consistency is expected. If researchers can reduce random noise in indicators (e.g., by designing unambiguous survey questions; by providing adequate time to respond to a survey; by providing respondents with a comfortable, distraction-free environment when implementing the survey; etc.) the reliability of an indicator should increase and the opportunity to measure the latent variable the researcher intended to measure (e.g., economic preferences) is more likely.

This article introduces a new approach to the valuation of economic preferences that facilitates the straightforward assessment of measurement reliability. In the following section, a brief history of reliability testing in contingent valuation is provided with reference to more substantial, earlier reviews of this literature by Jorgensen et al. [4] and Jorgensen [5]. Next, recent theoretical contributions to understanding the reliability of contingent values are discussed with reference to any significant strengths and limitations. Methods of testing the reliability of measured observations are discussed in Section 4 with particular reference to those methods directly applicable when multiple indicators are available. A new valuation approach – multiple-indicator contingent valuation (MCV) – is presented in Section 5 which enables the application of standard reliability assessment procedures because of its use of multiple-indicator measurement models. Readers are introduced to the approach in a step-by-step manner focusing on its implementation in applied settings. Section 6 briefly reviews the range the steps required to estimate the measurement model

and calculate the required reliability coefficients. Conclusions are provided in the final section of the article.

2. Reliability Assessment of Willingness-to-Pay (WTP) Responses in Contingent Valuation

Previous research on the reliability of measurements of WTP suffers from significant methodological flaws that have made their conclusions irrelevant to questions concerning random error [4, 6]. These errors continue to appear in recent research on the temporal stability of WTP responses [7-9]. The most frequently used means of estimating reliability in this literature has been the *Test-retest method* in which WTP is measured at two time points from the same sample of individuals. Consistency here is not limited to individuals providing the same response on two or more separate occasions. Rather, it refers to the consistency of the relative *rank-order* of individual responses on separate occasions. Responses on one occasion can demonstrate perfect reliability if their correlation with responses on a second occasion equals one.

The correlation between responses obtained at the two testing occasions constitutes the reliability coefficient. The test-retest correlation indexes the degree of reliability given that random sources of error variance are uncorrelated with systematic sources of variance. Correlations close to zero indicate low reliability and values close to ±1.0 reflect high reliability. In the Test-retest studies that are methodologically sound, most reliability coefficients are in the order of 0.50 to 0.70 suggesting moderate levels of reliability [4].

Since the review of reliability studies provided by Jorgensen et al. [4], some test-retest studies have continued to show modest levels of reliability [10]. Higher reliability ranging from 0.51 to 0.94 have been reported in several studies with relatively small samples of between 47 and 97 respondents [11-13]. All of these examples were conducted in health contexts valuing goods such as WTP to avoid 16 different health states [13]; WTP to obtain a vaccine [12]; and WTP for perfect health [11].

Measuring reliable preferences is a challenging task given that people have likely little experience in assigning a monetary value to changes in public goods, and in a manner that is consistent with the notion of economic preference [14]. Several researchers have noted the lability of these newly created preferences [15, 16] and contend that they are constructed in the very process of measurement [17]. Others have questioned whether the potential instability in WTP responses is the product of survey respondents having no attitude or opinion toward paying a certain amount for a public good [18, 19]. If the assumption that economic preferences for public goods are stable over time cannot be sustained, one must question the utility of test-retest reliability with its reliance on the temporal stability of a concept.

Venkatachalam [20] reported that a significant number of CV studies have been carried out without any concern for the reliability of the results. But, if preference indicators are not reliable measures, they have limited utility for hypothesis testing in validity trials [4]. This is because noise factors conceal causal effects on the dependent variable [21]. Demonstrating that individual WTP responses are reliable would provide some evidence that they are at least meaningful to respondents [19].

3. Applying the Concept of Reliability to Assess the Accuracy of Nonmarket Value Estimates

Mitchell and Carson [22] argued that researchers are required to "demonstrate that the individual willingness-to-pay amounts are not simply random responses" (p.213). To this end, they suggest that researchers show that their measure of WTP has a coefficient of determination of at least 0.15 when regressed on a small set of key independent variables. This exceptionally liberal cut-off is considerably lower than the conventional criterion of 0.70 in other areas of measurement in the social sciences. Researchers may seek to achieve the highest coefficient of determination possible but, as Mitchell and Carson state, "High R^2s for equations based on the kitchen sink approach in which explanatory power is maximized by the indiscriminate use of as many variables as possible, are not particularly meaningful especially in smaller samples…" (p.213). However, with this statement the authors recommend limiting the reliability analysis to maintain some semblance of validity. In the end, the suggestion restricts researchers from understanding either the reliability or the validity of their measurements.

Bishop and Boyle [23] note that assessments to date of the "accuracy" (i.e., the reliability and validity) of non-market values has been insufficient. They suggest that changing the way in which reliability is applied to non-market valuation research can lead to concomitant improvements in the quality of research. This seems a reasonable goal since it is apparent that understandings of reliability and its application to value measurement need to improve. However, the framework proposed by Bishop and Boyle focuses on the reliability of the *mean* value estimate calculated from the distribution of WTP/WTA values generated from a single valuation exercise. As a result, the assessment of reliability of the mean presupposes the existence of a distribution of mean values derived from several valuations of the exact same ecosystem service. The coefficient of reliability in this approach is the standard error of the grand mean, and "the larger the standard error, the less reliable the estimate" (p.562).

Bishop and Boyle [23] illustrate their conceptualisation of reliability in the following manner:

> "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." (p.560)

In this illustration, the archer represents different valuation exercises with the underlying logic being that WTP/WTA estimates from different valuations of the same public good should hit the same mark on the target. In standard approaches to reliability, based on classical test theory, focus rests with the unit of analysis, that is, the entity from which the measured response was obtained. In contingent valuation, that unit of analysis is the individual and not the sample from which the average WTP/WTA value is derived.

Furthermore, there is no requirement according to classical test theory to equate larger variance of the distribution of individual- or group-level responses with lower reliability. Rather, responses should be relatively free of random error variance. Measurements at the individual- or group-level can display significant variance due to systematic errors (i.e., measurement bias). Questions of systematic sources of variance are issues of validity and not reliability.

The approach to reliability proposed by Bishop and Boyle [23] requires access to several mean WTP/WTA estimates of the same ecosystem service or public good. Moreover, even studies focusing on the same good are likely to differ on a myriad of contextual factors including unmeasured characteristics of the participants, the time the valuations were undertaken, and characteristics of

the ecosystem providing the service being valued. This places significant limitations on the application of the approach since finding a set of comparable studies is likely to be difficult to achieve.

The above limitations notwithstanding, the biggest problem with the approach to reliability offered by Bishop and Boyle [23] is that it does not address the reliability of measurement. That is the degree of random error variance in the measured WTP/WTA responses of the individuals that provided them. Analysis at the mean level does not provide insights into the reliability of individual responses. Furthermore, as will become evident later in this article, it is unnecessary to dispense with classical notions of reliability for the sake of approaches that retreat from the unit of analysis at which the responses are generated (i.e., the individual level).

Jorgensen et al. [4] proposed that, by employing multiple indicator measurement models of WTP/WTA, reliability could be directly estimated for both the individual indicators and their combination. The authors recognized that the development of indicators that are responses to a bid offer that varies over subsamples of respondents (e.g., dichotomous choice WTP) is likely to be challenging. However, they suggested several avenues toward this end that might prove fruitful but, at the time of publication, required significant foundation research.

In the following section, a new approach to valuation that facilitates assessments of measurement accuracy is introduced. Like Jorgensen et al. [4] the method employs several indicators to measure the latent preference variable, but the indicators are simply standard WTP/WTA responses to questions that vary by bid amount. That is, significant development work to determine matters such as the best question-wording and bid vector design has already been conducted. This new approach is detailed in the remainder of the article.

4. Multiple-indicator Options for Reliability Assessment

Psychology has developed a general approach for measuring unobservable variables that relies on asking several survey questions that are designed to measure the same latent variable [24, 25]. The benefit of multiple-indicators exists because researchers have access to more than one indicator of the latent variable. Oftentimes, these indicators are combined to create a single measure with the idea that the random errors contained in each individual indicator will offset one another [26]. However, apart from any benefit that may or may not arise from scale development [27], the central benefit for current purposes is that multiple indicators enable reliability assessment in a straightforward manner. That is, by using several indicators of latent preferences no assumption of temporal stability is required and the reliability coefficient can be calculated from a single sample. No re-test sample is necessary, and stability over time of the latent concept is not assumed.

Multiple-indicator reliability assessment is based on the notion of *internal consistency* rather than stability over time. Reliability, therefore, is frequently assessed by examining the degree of correlation among multiple indicators, such that they are consistent with one another and measuring the same thing [28]. Larger correlations between items suggest better reliability because random error in measurements can attenuate correlations under certain conditions [29] or compromise the interpretation of hypothesis tests in other ways [30].

Cronbach's alpha (α) [31] is the most frequently used reliability coefficient for multiple indicator data and the Kuder-Richardson formulae (KR-20 and KR-21) are often used when the indicators are dichotomous [32]. However, there are a variety of ways to assess the reliability of multiple-

indicators and much debate about their relative strengths and weaknesses [33-35]. But there are practical strategies available to suit most kinds of data [36]. Interested readers are referred to this literature and to Jorgensen et al. [4] for an empirical example, as the remainder of this discussion will focus on describing the procedures required to undertake WTP measurement using MCV.

5. Multiple-indicator Contingent Valuation (MCV)

5.1 The Valuation Questions

CV surveys have employed a variety of WTP and Willingness-to-Accept (WTA) response formats: Open-ended, dichotomous choice; payment card, etc. MCV is an extension of the discrete response take-it-or-leave-it measurement strategy developed by [37]. Survey respondents are randomly assigned a value from a vector of prices (or bids) and then asked if they are willing to pay the price to gain or avoid the proposed change in the public good. MCV also employs a price vector but does not involve randomly assigning one value to each participant.[1] Rather, all prices are presented to each respondent as separate WTP questions, and the order by which these take-it-or-leave-it WTP questions are presented to participants is randomised. By randomly ordering the presentation of the questions for each individual, question-order effects such as assimilation and contrast effects are controlled [41].

The number of WTP questions needed to support estimation of a reliability coefficient depends upon the assessment approach taken. Cronbach's alpha for example is a simple model that assumes uncorrelated errors among indicators [42]. Alpha is based on the average correlation between indicators such that, in principle, it can be calculated with just two measures although the statistic would be based on just one correlation coefficient in that instance. More general confirmatory factor analysis measurement models require four or more indicators to support an assessment of measurement reliability concerning a single, latent variable [43]. This results in an *over-identified* measurement model for which all the parameters can be estimated.[2] An over-identified model enables an analysis of how well each indicator measures latent preferences and therefore the reliability of the WTP measures [44].

---

[1] The distribution of the selected prices for MCV can be guided by existing research on the subject (e.g., [38-40]).

[2] The parameter estimates of a model must be uniquely determined or *identified*. If a model is *underidentified*, any number of values for the parameter estimates could be derived that fit the data. If a model is identified, the following formula should hold:

$$t \leq s/2$$

where, t = number of parameters to be estimated.
s = number of non-redundant variances-covariances of the observed variables,

$$(p+q)(p+q+1).$$

P = number of y-variables.
q = number of x-variables.

A model is *just-identified* (or saturated) when $t = s/2$. Here, a unique solution can be estimated, but there is not enough degrees of freedom to subject the parameter estimates to a goodness-of-fit test. Because $df = 0$ for a just-identified model, its goodness-of-fit will appear to be perfect. A model is *over-identified* when $t < s/2$, although this is not a sufficient condition for identification.

While at least four WTP indicators are required for the measurement model to be over-identified, more indicators may be necessary to include in an MCV survey. In the take-it-or-leave-it approach, it is likely that responses to the lowest and highest bids will not be normally distributed (see, for example, Cooper and Loomis [39]). While measurement models can be estimated with non-normal data, researchers will need to screen the WTP responses for extreme departures from normality. In situations, where the standard deviations of WTP indicators reflect little variation in responses, researchers may need to restrict the reliability assessment to those indicators that meet whatever statistical assumptions their analysis requires. However, modern structural equation modelling software is capable of dealing with many types of data including dichotomous distributions, non-normal ordinal distributions, and non-normal latent variables [45].

5.2 Specifying the Intended Behaviour

Clearly specifying the intended behaviour in the valuation question is an important design consideration to reduce any perceived ambiguity and increase the likelihood that the intention will correspond with future behaviour [46, 47]. Presseau et al., recommend specifying the required behaviour using a framework encompassing five domains: Action, Actor, Context, Target, Time (AACTT). For example, the *action* relevant to contingent valuation is paying, voting, or accepting compensation. The *actors* applicable to a WTP behavioural intention are those persons who perform the action of paying, voting, or accepting compensation. Actors in CV surveys have frequently been individual respondents and households from a relevant population. The *target* in a valuation question is the focus of the behaviour and the object of valuation (i.e., the change in the public good). *Context* refers to the physical, emotional, or social conditions accompanying performance of the behaviour. In CV, the physical conditions and, to some extent the social and emotional conditions too, are likely linked to the contextual properties of the payment vehicle (e.g., entry charges to national parks, license fees for access to a fishing area, rates to local government, etc.). Finally, the *time* domain refers to when the behaviour will be performed. This might be the next time the individual visits a national park or pays for a fishing license. By clearly specifying the MCV behaviour in this way, ambiguity is reduced by making explicit who (actor) does what (action), when (time) and where (context), for what objective (target).

5.3 Selecting a Response Format

The response options for each WTP/WTA question can be either dichotomous (YES/NO) or polychotomous. Polychotomous response scales to WTP questions appear in the CV literature in studies of response uncertainty [38, 48] and commitment [49]. Anchors on the ordinal, polychotomous scale can vary depending upon how the WTP question is framed. In general, ordinal scales are preferred to dichotomous scales because the former indicate not only the valence of the response, but its intensity as well. Several WTP questions and response options are possible from a multi-indicator perspective and some of these are presented in Table 1.

All the examples in Table 1 include a price which is varied over subsamples. Moreover, all examples follow the AACTT structure discussed earlier, and the last question follows the referendum (voting) format while the remainder target payment behaviour. All these examples are amenable to a MCV approach because they comprise the same action, actor, context, target, and time.

The different wording of each question necessitates different anchors (or labels) to define the response options. While the anchors are somewhat intuitive given the wording of the questions, there is also a strong methodological literature to guide researchers if needed (e.g., [50-52]). One way of selecting anchors is to pre-test an open-ended version of the elicitation question on a pilot sample with frequently cited responses representing potential anchors in the MCV survey. Question-wording that elicits relatively high rates of "don't know" responses warrant further investigation and revision if they prove to be ambiguous for some reason.

While most of the example scales in Table 1 have a label for each response option, some authors recommend fewer labels [53] but there are advantages and disadvantages to both formats [52]. Pretesting alternative response formats can highlight any uncertainties for respondents. Perhaps more important than the labelling of response options is avoiding the use of "off-scale" options, for example, labelling the midpoint of an agree-disagree response scale as "don't know" and "unsure". These responses conflate a respondent's level of certainty in paying with their level of support for the change in the public good. Similarly, a midpoint of "no answer" might indicate that the respondent refuses to answer the question for some reason rather than signalling a response of mid-level agreement. In these cases, practitioners would be better served either leaving the midpoint unlabeled or labelling it as "neither agree nor disagree" [54].

Researchers may choose to vary the wording of the elicitation question within the MCV survey (as evident in Examples 1 to 4). These four questions are all examples of payment intentions and share the same AACTT components. While the price and wording can vary across elicitation questions, the response scale and anchors need to be consistent if researchers wish to estimate mean or median WTP values. But for the purpose of reliability assessment, there is no reason why the response scale labels could not also differ across elicitation questions. The need for consistency for the purpose of calculating mean and median WTP will become evident later in the article when the discussion shifts from one of reliability assessment to value estimation.

Table 1 WTP questions and response scales.

Elicitation Question	Possible Scale Anchors
1. I intend to pay $*x* more for my fishing licence next month to have Rainbow Smelt eradicated from Emerald Lake.	1. Strongly disagree 2. Disagree 3. Neither agree nor disagree 4. Agree 5. Strongly agree 1. Strongly disagree 2. Disagree 3. Somewhat disagree 4. Neither agree nor disagree 5. Somewhat agree 6. Agree 7. Strongly agree
2. Would you be willing to pay $*x* more for your fishing license next month if the money was used to eradicate Rainbow Smelt from Emerald Lake?	1. No 2. Yes 1. No

		2. Maybe
		3. Yes
		1. Definitely no
		2. Probably no
		3. Maybe yes, maybe no
		4. Probably yes
		5. Definitely yes
		1. Very unlikely
3.	How likely is it that you would pay an additional $*x* next month for your fishing license if the money was used to eradicate Rainbow Smelt from Emerald Lake?	2. Fairly unlikely
		3. Neither likely nor unlikely
		4. Fairly likely
		5. Very likely
		1. Very unwilling
4.	How willing are you to pay an additional $*x* next month for your fishing licence if the money was used to eradicate Rainbow Smelt from Emerald Lake?	2. Fairly unwilling
		3. Neither willing nor unwilling
		4. Fairly willing
		5. Very willing
		1. Definitely not support it
5.	If I was asked to vote in a referendum on a proposal to increase the cost of a fishing licence next month by an additional $*x* so that the money raised could be used to eradicate Rainbow Smelt from Emerald Lake, I would …	2. Probably not support it
		3. Might or might not support it
		4. Probably support it
		5. Definitely support it
		1. Reject it
		7. Accept it

In sum, researchers employing MCV might use just one type of WTP question and response scale but repeat it with only the price bid changing. Alternatively, different types of WTP questions might be used *and* vary the price offered. However researchers decide to structure the valuation question and what response scales they choose, there should be as many questions as there are bids in the price vector, and at least four bids, questions, and responses.

6. Model-based Reliability Estimation

As noted earlier in the discussion, there are several methods to estimate the reliability of multiple indicators, and some of these depend upon the properties of the response data (e.g., its metric; distribution, etc.). To illustrate, a confirmatory factor analysis model-based approach will be described (see also [4]). Model-based approaches to reliability are superior to simpler approaches such as Cronbach's α because the relationship between the theoretical latent variable and its indicators is made explicit [33]. Restrictive assumptions such as uncorrelated errors among indicators can be formerly examined and subjected to hypothesis testing. In this sense, reliability assessment is a by-product of a specific measurement model rather than a "stand-alone" enterprise [34].

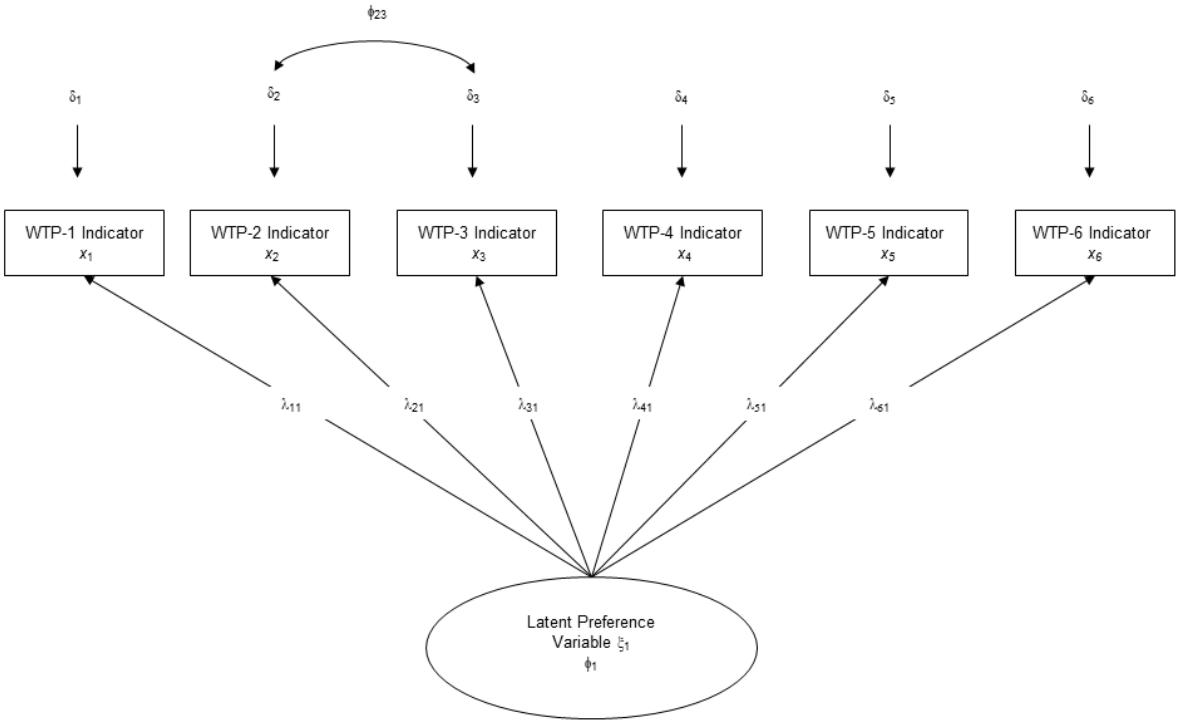

Figure 1 Unidimensional measurement model with congeneric indicators.

A simple measurement model is presented in Figure 1 that shows a single latent preference variable (ξ_j) and six endogenous observed WTP indicators (x_i). Slope parameters (λ_{ij}) link the observed variables to the latent variable and are represented in the lambda (Λ) matrix. The linear regression coefficients also include a vector of intercepts τ_i. The residual variances (δ_i) of each indicator are contained in the theta matrix (Θ) and may comprise both random and systematic error variance not explained by the latent variable. Shared systematic residual variance accounts for correlations between the errors of two or more indicators which are represented as off-diagonal elements (θ_{ij}) in the theta matrix. Therefore, observed indicators are represented as a linear function of a latent variable, an intercept, and a stochastic error term:

$$x_i = \tau_i + \lambda_{ij}\xi_j + \delta_i$$

The parameters of the measurement model can be estimated using a variety of commercially available software. The results provide all the information necessary for calculating several reliability coefficients. The reliability of individual WTP indicators is given by the coefficient of determination from the regressions on the latent variable [44]:

$$R^2{}_{xi} = \lambda^2{}_{ij}\phi_i/\left(\lambda^2{}_{ij}\phi_i + \delta_i\right)$$

where ϕ_i is the variance of item *i*.

To obtain a reliability coefficient for the combination of WTP indicators in Figure 1, researchers can calculate a reliability coefficient that has its origins in the work of [55-57]. The coefficient is variously referred to as *construct reliability, composite reliability*, or *coefficient omega* [58] and is calculated with the following formula:

$$\rho c = \left(\sum \lambda_{ij}\right)^2 / \left(\sum \lambda_{ij}\right)^2 + \sum \delta_j$$

In the formula above $(\sum \lambda_{ij})^2$ is the true-score variance and $\sum \delta_j$ is the error variance. All the information necessary to calculate ρc is generated from the estimation of the measurement model. If this model-based coefficient proves that the indicators are sufficiently reliability (i.e., ρc > 0.70 by convention) then researchers can proceed to the estimation of mean WTP. Reliability assessed within a MCV framework means that researchers know whether their measurement data is reliable, rather than having to generalise from test-retest studies undertaken in sometimes completely different contexts.

6.1 Estimating the Measurement Model

The model described in Figure 1 was estimated using the data simulation function in Mplus 8.8 [59]. The range of reliability estimates available from the environmental economics literature were used to set the indicator loadings and error variances. The review by Jorgensen et al. [4] provides reliability coefficients ranging from 0.30 to 0.95. The following values were selected for the purpose of illustrating how reliability coefficients can be estimated from a measurement model like the one in Figure 1: 0.30, 0.40, 0.59, 0.71, 0.79, 0.95.

The model estimates are shown in Figure 2. Several goodness-of-fit indices are provided upon which to assess the adequacy of the model. For example, the model chi-square (degrees-of-freedom) is 7.97(8) suggesting that the variance-covariance matrix estimated from the model is a close fit to the actual sample variance-covariance matrix. The reliability coefficient for each indicator is the R^2 value which were set to values consistent with Jorgensen et al. [4]. These values reflect the value of the standardised loadings (λ_{ij}) and error variances (δ_j) for their respective indicators. The composite reliability (ρc) can be calculated from the loadings and error variances using the equation from the previous section. This equation produced a reliability coefficient of 0.90.

This ρc estimate is marginally larger than Cronbach's alpha (α = 0.89) calculated with SPSS 28 [60] from the correlation matrix simulated from the parameters of the measurement model. Part of the output information provided in the estimation of alpha is the identification of indicators responsible for low values. In this illustration, the output suggested that the removal of WTP$198 would serve to increase alpha by a trivial 0.01 to a value of 0.90. That is, but two different procedures, the reliability estimates, if not identical, are extremely close.

As the simulated modelling has illustrated, reliability coefficients can be easily calculated for individual indicators and for a set of indicators. The latent structure of the WTP responses can also be subject to hypothesis tests to identify the number of latent preference variables required to explain observed responses. This is done by estimating a range of models varying in the number of latent variables posited and the pattern of indicator loadings on those latent variables (see, for example, Jorgensen and Stedman, [61]). As noted by Jorgensen et al. ([4], p.50), the model-based approach to reliability (and validity):

"…requires CV practitioners to think more about the constructs that they wish to measure. Specifying and testing particular models informs questions of reliability and validity at the same time."

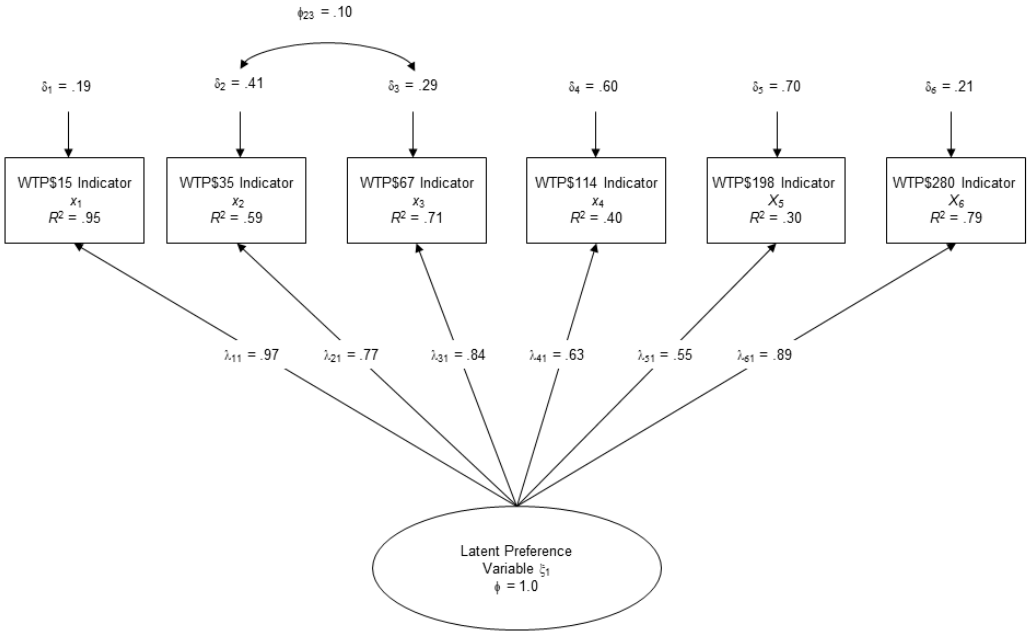

Figure 2 Results of the simulated model estimation.

While the parameter estimates in Figure 2 were obtained using simulated data, interested readers can generate real-world data by simply following the procedures and examples presented in earlier sections of this article.

7. Estimating Mean WTP

7.1 Random Selection of WTP Responses

MCV data can be re-configured to be identical to standard take-it-or-leave-it data which are easily used to estimate mean and median WTP values. To establish a take-it-or-leave-it basis for estimating mean WTP, a randomly assigned bid and associated WTP response are required for every case in the sample. If the researcher has a vector of prices that contains five bids, for example, and there are five corresponding WTP questions randomised for each individual, then only one bid and response from each participant is required to derive a mean WTP value (see Figure 3).

To create a subsample in which all participants have just one bid and one WTP response, researchers can randomly select an individual's WTP response from the five questions he or she was presented with during surveying. The process of randomly selecting WTP responses from each individual needs to be consistent with the size of the subsamples for which the bids were varied. In standard CV it is usually the case that the same number of participants receive each bid from the price vector.

This process leads to a data set in which each participant is associated with the response to just one of the five WTP questions they were initially asked. For the sample, there are an equal number of participants who received each bid since it is this value that was the focus of the random selection process and not the WTP response itself. The result is a dataset that has the same form as might be produced from a standard take-it-or-leave-it CV survey.

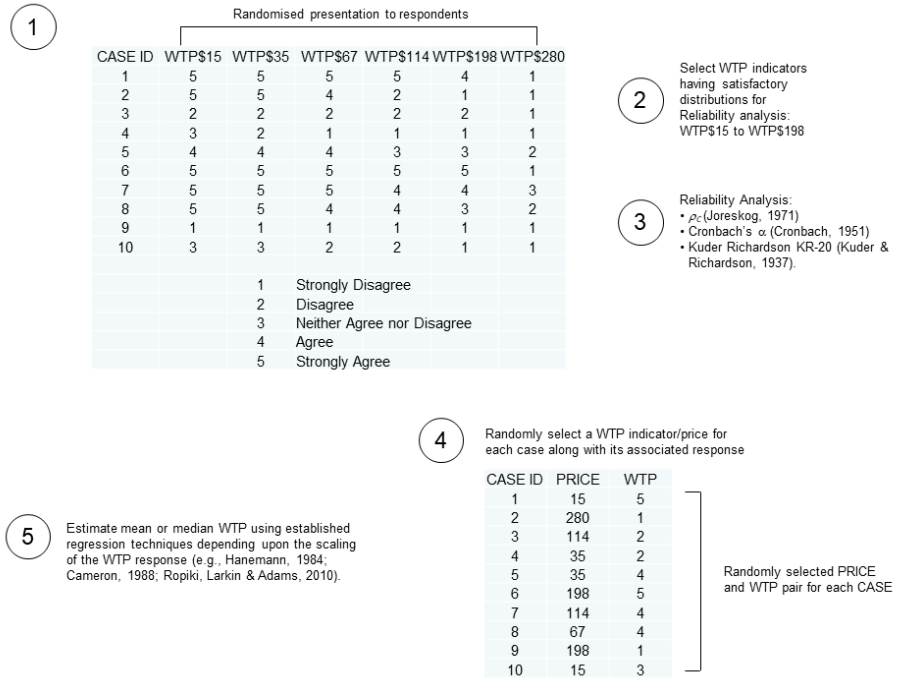

Figure 3 Data transformation process.

For the estimation of mean WTP from dichotomous response data, the regression techniques are well documented and researched [62, 63]. Researchers who have used polychotomous response scaling have also adopted a variety of regression approaches depending upon their objectives [38, 48, 49, 64]. However, among these studies, recoding the polychotomous categories into a dichotomous response variable frequently occurs. As noted previously in the discussion, this strategy, while preserving the valence of the responses, loses information about the intensity of the response. Therefore, instead of rescaling the WTP variable, researchers might follow Obeng and Aguilar [65] who use ordered logistic regression to estimate mean WTP from their polychotomous responses. However, adopting this approach requires that the data satisfies the assumption of parallel regressions across the levels of the polychotomous WTP variable [66].

7.2 Obtaining a Distribution of WTP Sample Means

MCV offers researchers not only a single sample way of estimating the reliability of WTP measurements, but it enables the estimation of a probability distribution of the WTP mean simply by repeatedly resampling from the data. This latter benefit arises from the randomised, multiple indicator innovation which is characteristic of MCV. Recall from Figure 3 that one WTP response was randomly selected (within price levels) from a possible six responses. By repeating this process and obtaining different responses on each occasion creates a distribution of WTP means to be calculated from a single MCV dataset. If data from six WTP questions were elicited from 100 respondents, the

total number of combinations and permutations approaches one million. From a sufficient number of observations and participants, the distribution of WTP means can provide a standard error of the grand mean thereby providing researchers with an estimate of the uncertainty associated with this WTP estimate.

8. Reliability and Validity

It is crucial to establish the reliability of a variable prior to undertaking validity tests because reliability is a necessary but not sufficient condition for validity [3]. A variable that is not reliable cannot be valid. Failed validity tests can therefore be due to unreliability rather than to any theoretical explanation underpinning the validity test [4, 67].

MCV can establish reliability using a small set of indicators and statistical models that can accommodate different types of data including non-normal distributions. Given an adequate reliability coefficient, researchers might proceed to assessing validity prior to the estimation of mean and median WTP values. On the occasions that those validity tests fail, researchers will have confidence in discounting poor measurement reliability as an explanation.

The nature of validity tests in CV will ultimately depend upon how firmly researchers are committed to the logical requirements of rationality as a validity criterion. The challenge of determining trade-off rates using hypothetical markets is considerable and the reliable and valid measurement of well-defined and continuous preferences is essential. However, rational choice theory is not a necessary basis of validity in MCV. Rather, practitioners are free to scrutinise their data against whatever theory they choose. The reliable and valid measurement of behavioural intentions (e.g., WTP) supports accurate predictions of *actual* behaviour but not at the cost of the theoretical axioms that describe economic preferences.

When viewed as a behavioural intentions, factors that contribute to the reliability of WTP (e.g., protest responses) but are dismissed from a neoclassical economic perspective (because they are assumed to arise from non-compensatory preferences) are, nonetheless, completely valid motivations of actual behaviour [68, 69]. Furthermore, there is a well-established literature identifying the factors and conditions that influence the likelihood that intentions will translate into actual performance of the behaviour [46, 70-72]. Researchers might ask individuals if they are willing to pay a given amount of money for a fishing license next month if the funds raised were used to eradicate rainbow smelt from Emerald Lake with the validity criterion being actual payment under the conditions explicit in the question (i.e., action, actor, context, target, and time). This behavioural criterion changes researchers' determinations of the valid sources of variance in WTP responses, such that, some types of bias once considered reliable but not valid influences in CV may be re-considered as both reliable *and* valid.

9. Conclusion

Contingent valuation research, and stated preference research in general, has enthusiastically pursued the identification and avoidance of biases in preference measurement and afforded comparably little effort in establishing the extent to which the method can provide consistent individual responses. Past research on reliability has been fraught with methodological and conceptual limitations that have rendered their interpretation difficult. Furthermore, the demands of drawing a re-test sample (of the same individuals in the initial sample) in tests of the reliability of

individual WTP responses limits it proliferation in the literature in addition to its assumption that labile preferences are stable over time.

A new approach to measuring preferences – Multiple-indicator Contingent Valuation (MCV) – enables the assessment of reliability in several ways and with the same sample from which benefit estimates are derived. Assessment of reliability from a single sample alleviates the need to undertake more resource intensive methodologies such as test-retest reliability. This innovation will support an increase in reliability studies since every application of MCV can produce at least one estimate of measurement reliability. With this renewed interest in reliability assessment, the field can substantially improve its understanding of the conditions that influence the reliability of preference indicators because independent variables can be manipulated in the same research design that generates the reliability coefficient. Whatever way the field of non-market valuation progresses, it is imperative that evidence of the reliability of WTP responses is properly generated and that these results are convincingly scrutinised for policymakers to comprehend their meaning and implications in the contexts where they are to be applied.

Author Contributions

The author did all the research work of this study.

Competing Interests

The author has declared that no competing interests exist

References

1. Bollen KA. Latent variables in psychology and the social sciences. Annu Rev Psychol. 2002; 53: 605-634.
2. Costner HL. Theory, deduction, and rules of correspondence. Am J Sociol. 1969; 75: 245-263.
3. Meyer P. Understanding measurement: Reliability. New York: Oxford University Press; 2010.
4. Jorgensen BS, Syme GJ, Smith LM, Bishop BJ. Random error in willingness to pay measurement: A multiple indicators, latent variable approach to the reliability of contingent values. J Econ Psychol. 2004; 25: 41-59.
5. Jorgensen BS. The determinants of assigned value: A social psychological approach to welfare measurement. Perth: Curtin University; 1996.
6. Jorgensen BS. Perceived justice and the economic valuation of the environment: A role for fair decision-making procedures. In: Towards an environment research agenda: A second selection of papers. London: Palgrave Macmillan; 2003. pp. 146-161.
7. He J, Zhang B. Current air pollution and willingness to pay for better air quality: Revisiting the temporal reliability of the contingent valuation method. Environ Resour Econ. 2021; 79: 135-168.
8. Perni Á, Barreiro-Hurlé J, Martínez-Paz JM. Contingent valuation estimates for environmental goods: Validity and reliability. Ecol Econ. 2021; 189: 107144.
9. Williams G. The temporal stability of WTP estimates for the emissions reduction using the contingent valuation survey in Queensland, Australia. J Environ Econ Policy. 2022. Doi: 10.1080/21606544.2022.2149628.

10. Onwujekwe O, Fox-Rushby J, Hanson K. Inter-rater and test–retest reliability of three contingent valuation question formats in south-east Nigeria. Health Econ. 2005; 14: 529-536.
11. Mavrodi AG, Chatzopoulos SA, Aletras VH. Examining willingness-to-pay and zero valuations for a health improvement with logistic regression. Inquiry. 2021; 58: 00469580211028102.
12. Shiell A, Hawe P. Test-retest reliability of willingness to pay. Eur J Health Econ. 2006; 7: 173-178.
13. Smith RD. The reliability of willingness to pay for changes in health status. Appl Health Econ Health Policy. 2004; 3: 35-38.
14. Kahneman D. Comments on the contingent valuation method. In: Valuing environmental goods: An assessment of the contingent valuation method. Totowa: Roweman and Allanheld; 1986. pp. 185-193.
15. Kahneman D, Ritov I, Schkade D, Sherman SJ, Varian HR. Economic preferences or attitude expressions? An analysis of dollar responses to public issues. In: Elicitation of preferences. Dordrecht: Springer; 1999. pp. 203-242.
16. Kahneman D, Ritov I, Jacowitz KE, Grant P. Stated willingness to pay for public goods: A psychological perspective. Psychol Sci. 1993; 4: 310-315.
17. Lichtenstein S, Slovic P. The construction of preference. New York: Cambridge University Press; 2006.
18. Jorgensen BS, Syme GJ, Nancarrow BE. The role of uncertainty in the relationship between fairness evaluations and willingness to pay. Ecol Econ. 2006; 56: 104-124.
19. Schuman H. The sensitivity of CV outcomes to CV survey methods. In: The contingent valuation of environmental resources: Methodological issues and research needs. Cheltenham: Edward Elgar Publishing; 1996. pp. 75-96.
20. Venkatachalam L. The contingent valuation method: A review. Environ Impact Assess Rev. 2004; 24: 89-124.
21. Shadish WR, Cook TD, Campbell DT. Experimental and quasi-experimental designs for generalized causal inference. Boston: Houghton, Mifflin and Company; 2002.
22. Mitchell RC, Carson RT. Using surveys to value public goods: The contingent valuation method. Washington: Rff Press; 2013.
23. Bishop RC, Boyle KJ. Reliability and validity in nonmarket valuation. Environ Resour Econ. 2019; 72: 559-582.
24. Curtis RF, Jackson EF. Multiple indicators in survey research. Am J Sociol. 1962; 68: 195-204.
25. Sullivan JL, Feldman S. Multiple indicators: An introduction. Thousand Oaks: SAGE; 1979.
26. Nunnally JC, Bernstein IH. Psychometric theory. 3rd ed. McGraw-Hill series in psychology. New York: McGraw-Hill; 1994.
27. Drolet AL, Morrison DG. Do we really need multiple-item measures in service research? J Serv Res. 2001; 3: 196-204.
28. Streiner DL. Starting at the beginning: An introduction to coefficient alpha and internal consistency. J Pers Assess. 2003; 80: 99-103.
29. Rigdon EE. Demonstrating the effects of unmodeled random measurement error. Struct Equ Modeling. 1994; 1: 375-380.
30. Loken E, Gelman A. Measurement error and the replication crisis. Science. 2017; 355: 584-585.
31. Cronbach LJ. Coefficient alpha and the internal structure of tests. Psychometrika. 1951; 16: 297-334.

32. Kuder GF, Richardson MW. The theory of the estimation of test reliability. Psychometrika. 1937; 2: 151-160.
33. Bentler PM. Alpha, dimension-free, and model-based internal consistency reliability. Psychometrika. 2009; 74: 137-143.
34. Sijtsma K. On the use, the misuse, and the very limited usefulness of Cronbach's alpha. Psychometrika. 2009; 74: 107-120.
35. Viladrich C, Angulo-Brunet A, Doval E. A journey around alpha and omega to estimate internal consistency reliability. Ann Psychol. 2017; 33: 755-782.
36. Trizano-Hermosilla I, Alvarado JM. Best alternatives to Cronbach's alpha reliability in realistic conditions: Congeneric and asymmetrical measurements. Front Psychol. 2016; 7: 769.
37. Bishop RC, Heberlein TA. Measuring values of extramarket goods: Are indirect measures biased? Am J Agric Econ. 1979; 61: 926-930.
38. Alberini A, Boyle K, Welsh M. Analysis of contingent valuation data with multiple bids and response options allowing respondents to express uncertainty. J Environ Econ Manage. 2003; 45: 40-62.
39. Cooper JC, Hanemann M, Signorello G. One-and-one-half-bound dichotomous choice contingent valuation. Rev Econ Stat. 2002; 84: 742-750.
40. Kanninen BJ, Kriström B. Sensitivity of willingness-to-pay estimates to bid design in dichotomous choice valuation models: Comment. Land Econ. 1993; 69: 199-202.
41. Rasinski KA, Lee L, Krishnamurty P. Question order effects. In: APA handbook of research methods in psychology, vol 1: Foundations, planning, measures, and psychometrics. Washington: American Psychological Association; 2012. pp. 229-248.
42. Miller MB. Coefficient alpha: A basic introduction from the perspectives of classical test theory and structural equation modeling. Struct Equ Modeling. 1995; 2: 255-273.
43. Reilly T, O'BRIEN RM. Identification of confirmatory factor analysis models of arbitrary complexity: The side-by-side rule. Sociol Methods Res. 1996; 24: 473-491.
44. Bollen KA. Structural equations with latent variables. Toronto: John Wiley & Sons; 1989.
45. Wall MM, Guo J, Amemiya Y. Mixture factor analysis for approximating a nonnormally distributed continuous latent factor with continuous and dichotomous observed variables. Multivariate Behav Res. 2012; 47: 276-313.
46. Fishbein M, Ajzen I. Predicting and changing behavior: The reasoned action approach. New York: Psychology press; 2010.
47. Presseau J, McCleary N, Lorencatto F, Patey AM, Grimshaw JM, Francis JJ. Action, actor, context, target, time (AACTT): A framework for specifying behaviour. Implement Sci. 2019; 14: 102.
48. Ready RC, Whitehead JC, Blomquist GC. Contingent valuation when respondents are ambivalent. J Environ Econ Manage. 1995; 29: 181-196.
49. Ropicki AJ, Larkin SL, Adams CM. Seafood substitution and mislabeling: WTP for a locally caught grouper labeling program in Florida. Mar Resour Econ. 2010; 25: 77-93.
50. Converse JM, Presser S. Survey questions: Handcrafting the standardized questionnaire, survey questions: Handcrafting the standardized questionnaire. Thousand Oaks: SAGE Publications, Inc.; 1986.
51. Fink A. The survey handbook. 2nd ed. Thousand Oaks: SAGE Publications, Inc.; 2003.
52. Robinson SB, Leonard KF. Designing quality survey questions. 1st ed. Los Angeles: SAGE Publications, Inc.; 2018.

53. Darbyshire P, McDonald H. Choosing response scale labels and length: Guidance for researchers and clients. Australas J Mark Res. 2004; 12: 17-26.
54. Shishido K, Iwai N, Yasuda T. Designing response categories of agreement scales for cross-national surveys in east Asia: The approach of the Japanese General Social Surveys. Jpn J Soc. 2009; 18: 97-111.
55. Jöreskog KG. Statistical analysis of sets of congeneric tests. Psychometrika. 1971; 36: 109-133.
56. McDonald RP. The theoretical foundations of principal factor analysis, canonical factor analysis, and alpha factor analysis. Br J Math Stat Psychol. 1970; 23: 1-21.
57. Werts CE, Linn RL, Jöreskog KG. Intraclass reliability estimates: Testing structural assumptions. Educ Psychol Meas. 1974; 34: 25-33.
58. Cho E. Making reliability reliable: A systematic approach to reliability coefficients. Organ Res Methods. 2016; 19: 651-682.
59. Muthén LK, Muthén BO. Mplus user's guide. 8th ed. Los Angeles: Muthen & Muthen; 1998.
60. IBM Corp. IBM SPSS statistics for windows, version 28.0. Armonk: IBM Corp; 2021.
61. Jorgensen BS, Stedman RC. Sense of place as an attitude: Lakeshore owners attitudes toward their properties. J Environ Psychol. 2001; 21: 233-248.
62. Cameron TA. A new paradigm for valuing non-market goods using referendum data: Maximum likelihood estimation by censored logistic regression. J Environ Econ Manage. 1988; 15: 355-379.
63. Hanemann WM. Welfare evaluations in contingent valuation experiments with discrete responses. Am J Agric Econ. 1984; 66: 332-341.
64. Whitehead JC, Huang JC, Blomquist GC, Ready RC. Construct validity of dichotomous and polychotomous choice contingent valuation questions. Environ Resour Econ. 1998; 11: 107-116.
65. Obeng EA, Aguilar FX. Value orientation and payment for ecosystem services: Perceived detrimental consequences lead to willingness-to-pay for ecosystem services. J Environ Econ Manage. 2018; 206: 458-471.
66. Brant R. Assessing proportionality in the proportional odds model for ordinal logistic regression. Biometrics. 1990; 46: 1171-1178.
67. Jorgensen BS, Wilson MA, Heberlein TA. Fairness in the contingent valuation of environmental public goods: Attitude toward paying for environmental improvements at two levels of scope. Ecol Econ. 2001; 36: 133-148.
68. Jorgensen BS, Syme GJ, Bishop BJ, Nancarrow BE. Protest responses in contingent valuation. Environ Resour Econ. 1999; 14: 131-150.
69. Jorgensen BS, Syme GJ, Lindsey G. Discussion and closure: Market models, protest bids, and outliers in contingent valuation. J Water Resour Plan Manag. 1995; 121: 400-402.
70. Sheeran P. Intention—behavior relations: A conceptual and empirical review. Eur Rev Soc Psychol. 2002; 12: 1-36.
71. Sheeran P, Webb TL. The intention–Behavior gap. Soc Personal Psychol Compass. 2016; 10: 503-518.
72. Jorgensen BS, Boulet M, Hoek AC. A Level-of-analysis issue in resource consumption and environmental behavior research: A theoretical and empirical contradiction. J Environ Manage. 2020; 260: 110154.