Social Science & Medicine 232 (2019) 238–261

Contents lists available at ScienceDirect

Social Science & Medicine

journal homepage: www.elsevier.com/locate/socscimed

Review article

# The criterion validity of willingness to pay methods: A systematic review and meta-analysis of the evidence

Lucy Kanya[a,*,1], Sabina Sanghera[a,b], Alex Lewin[c,2], Julia Fox-Rushby[a,3]

[a] Health Economics Research Group, Brunel University, Kingston Lane, Uxbridge, UB8 3PH, England, UK
[b] Bristol Medical School (Population Health Sciences), Bristol University, 1-5 Whiteladies Road, Bristol, England, UK
[c] Department of Mathematics and Statistics, Brunel University, Kingston Lane, Uxbridge, UB8 3PH, England, UK

ARTICLE INFO

*Keywords:*
Contingent valuation
Willingness to pay
External validity
Criterion validity
Hypothetical values
Simulated market experiments
Systematic review
Meta-analysis

ABSTRACT

*Background:* The contingent valuation (CV) method is used to estimate the willingness to pay (WTP) for services and products to inform cost benefit analyses (CBA). A long-standing criticism that stated WTP estimates may be poor indicators of actual WTP, calls into question their validity and the use of such estimates for welfare evaluation, especially in the health sector. Available evidence on the validity of CV studies so far is inconclusive. We systematically reviewed the literature to (1) synthesize the evidence on the criterion validity of WTP/willingness to accept (WTA), (2) undertake a meta-analysis, pooling evidence on the extent of variation between stated and actual WTP values and, (3) explore the reasons for the variation.

*Methods:* Eight electronic databases were searched, along with citations and reference reviews. 50 papers detailing 159 comparisons were identified and reviewed using a standard proforma. Two reviewers each were involved in the paper selection, review and data extraction. Meta-analysis was conducted using random effects models for ratios of means and percentage differences separately. Meta-bias was investigated using funnel plots.

*Results:* Hypothetical WTP was on average 3.2 times greater than actual WTP, with a range of 0.7–11.8 and 5.7 (0.0–13.6) for ratios of means and percentage differences respectively. However, key methodological differences between surveys of hypothetical and actual values were found. In the meta-analysis, high levels of heterogeneity existed. The overall effect size for mean summaries was 1.79 (1.56–2.04) and 2.37 (1.93–2.80) for percent summaries. Regression analyses identified mixed results on the influence of the different experimental protocols on the variation between stated and actual WTP values. Results indicating publication bias did not account for differences in study design.

*Conclusions:* The evidence on the criterion validity for CV studies is more mixed than authors are representing because substantial differences in study design between hypothetical and actual WTP/WTA surveys are not accounted for.

## 1. Introduction

Cost-benefit analysis (CBA) of public investments requires measurement of aggregate WTP (Slothuus, 2000). The CV method allows the assignment of a monetary value to the benefits attached to a public good or service for comparison with its costs (Mitchell and Carson, 1989). In this way, the method enables the estimation of economic value for a wide range of commodities not traded in markets (Slothuus et al., 2002). Surveys or interviews are used to elicit people's preferences and monetary valuation for goods or services by asking about their WTP or WTA (Mitchell and Carson, 1989). By assuming a utility-theoretic model of consumer preferences, utility is maximised through the consumption of quantities of a good (or service) regarded as a "good" (*ceteris paribus*) (Klose, 2003). On the other hand, when a good (or service) is regarded as "bad", utility is maximised by consuming (purchasing) less of it. The maximum WTP or minimum willingness to accept (WTA) values for provision (loss) of goods or improvements (reductions) in services signal the individuals' valuation of

* Corresponding author. Department of Health Policy, London School of Economics and Political Science, Houghton Street, London, WC2A 2AE, UK.
*E-mail addresses:* L.Kanya@lse.ac.uk (L. Kanya), Sabina.sanghera@bristol.ac.uk (S. Sanghera), Alex.Lewin@lshtm.ac.uk (A. Lewin), Julia.Fox-Rushby@kcl.ac.uk (J. Fox-Rushby).
[1] Present address: Department of Health Policy, London School of Economics and Political Science, Houghton Street, London, WC2A 2AE, UK.
[2] Present address: Department of Medical Statistics, Faculty of Epidemiology and Population Health, LSHTM, Keppel Street, London, WC1E 7HT, UK.
[3] Present address: School of Population Health & Environmental Sciences, Faculty of Life Sciences & Medicine, Guys Campus, Kings College London, SE1 9RT, UK.

https://doi.org/10.1016/j.socscimed.2019.04.015
Received 22 November 2018; Received in revised form 11 March 2019; Accepted 14 April 2019
Available online 25 April 2019

the situation with or without the good/service (Mitchell and Carson, 1989). The WTP (or WTA) values are elicited contingent on a market existing for the valuation goods.

Unlike other preference elicitation methods such as travel cost (TCM), hedonic pricing method (HP), conjoint analysis and averting expenditures or averting behaviour, CV can be used to estimate both use and non-use values. CV therefore represents the most promising approach yet developed for determining the public's WTP especially for public goods. The CV is the most widely used yet controversial of methods to value non-marketed goods (Munro, 2009). While the CV method has been widely used in the environmental and transport sectors, it has been less frequently applied in the health sector. Significant concerns about the use of the method focus on the validity of estimates with critics arguing that hypothetical WTP values do not accurately reflect actual values (Loomis et al. 1996, 1997, 2009; Blumenschein et al., 2001; Blumenschein et al. 1998). This difference between hypothetical and actual values has been defined as hypothetical bias. There are also concerns about the potential for other biases relating to both the researcher (e.g. design bias) and the survey respondents (e.g. strategic bias) when using the CV method (Brown and Taylor, 2000; Champ et al., 1997; Cummings et al. 1995, 1997).

The extent to which hypothetical estimates of mean WTP reflect true values can be assessed using: (i) content validity, which reflects the extent to which an empirical measurement adequately reflects a specific domain of content (Carmines and Zeller, 1979). In WTP this is reflected in whether the framing of the CV questions for the good being valued is appropriate; (ii) construct validity, which concerns the correspondence between a measure and other measures of the same construct, and the degree to which the findings of a study are consistent with theoretical expectations. For example, construct validity may be assessed by measuring the convergence between values generated using a CV study and other preference elicitation measures such as the TCM. Theoretical relationships may also be tested by comparing mean WTP values of different conditions for which theory suggests different values (Hanley and Splash, 1993; Mitchell and Carson, 1989); and (iii) criterion validity, which is defined as the correlation of a scale with another measure of the trait, ideally a gold standard which has been used and accepted in the field. Criterion validity is assessed through either concurrent validity in which a new measure is correlated with an existing gold standard with data for both collected at the same time; or through predictive validity in which the new criterion is not yet available as at the time of data collection.

Criterion validity has the greatest potential for offering a definitive test of a measure's WTP validity (Mitchell and Carson, 1989). Actual market prices have been taken as an important criterion in CV studies. However, market prices are rarely available for public and quasi-public goods that generate significant non-use values, and therefore often no ideal criterion validity tests are available (Brown et al., 1996). In this absence, experimental (simulated) markets, in which the outcomes of hypothetical CV markets are compared with outcomes for identical markets in which the same goods are bought or sold, have been used. The actual (real) payment values generated from the simulated market experiments are compared against hypothetical values to evaluate criterion validity (Mitchell and Carson, 1989).

To date, 6 reviews (four of which include meta-analysis of values obtained) of criterion validity have been conducted (Carson et al., 1996; Harrison and Rutström, 2008; Liljas and Blumenschein, 2000; List and Gallet, 2001; Little and Berrens, 2003; Murphy et al., 2004) across different sectors. The evidence from these reviews further confirms the presence of hypothetical bias in CV-WTP studies. The effects of different experimental protocols on hypothetical bias have been investigated with mixed results. For example, the variety of elicitation formats, subject pools, study designs (whether within-group or between group), whether the welfare measure is WTP or WTA and the type of good (private or public) have been identified as potential drivers of hypothetical bias. However, the effect of these on hypothetical bias is mixed across the reviews. The last review was conducted more than a decade ago (Little and Berrens, 2003). The review by Harrison and Rutström (2008) was conducted in 1999 but was not published until 2008. Only two criterion validity assessments of health goods were included in the synthesised evidence to this date (Bhatia and Fox-Rushby, 2003; Blumenschein et al., 2001).

In a review of literature in 1998, Smith argued that data for criterion validity assessments – the 'gold standard' – was not available (Smith et al., 1999). However, since this date, both data on this gold standard and its use in the health sector have developed substantially, with some authors arguing that "the potential for survey instruments to provide valid estimates of WTP has been proven" (Donaldson and Shackley, 2002). However, there remains great concern about whether hypothetical values provide correct estimates of actual WTP and the evidence appears to be mixed (Munro, 2009; Loomis et al. 1996, 1997; Blumenschein et al., 2001). With more recent studies comparing stated and actual values performed since the last review, meta-analyses of the summary values will hopefully show consistent results regarding the magnitude of hypothetical bias.

This paper presents a narrative and quantitative systematic review and meta-analysis assessing the criterion validity of WTP methods. The review seeks to provide current evidence across the sectors on the criterion validity of WTP methods. This review differs from previous systematic reviews and meta-analyses of criterion validity assessments in two ways; we include (1) only criterion validity assessments which include direct WTP elicitation methods only in both the hypothetical and actual surveys and (2) only studies which report empirical WTP or WTA values. These criteria justify the broad search which identified some of the studies included in previous studies. An updated review will potentially highlight improvements in both the conduct and analysis of criterion validity assessments and may derive important methodological findings regarding WTP CV methods.

## 2. Methods

The review follows the PRISMA guidance on methods for conducting and reporting of systematic reviews (Moher et al., 2009).

### 2.1. Literature search strategy

Eight electronic databases (EconLit, TRID, MEDLINE, Embase, Web of Science, Psychinfo, CRD and CINAHL Plus) were searched from their inception to September 2016. The search terms were identified from previous systematic reviews (Carson et al., 1996; Harrison and Rutström, 2008; Liljas and Blumenschein, 2000; List and Gallet, 2001; Little and Berrens, 2003; Murphy et al., 2004). Valuation terms (WTP, WTA, CV, hypothetical value, hypothetical market, indirect, stated preference, stated value, actual market, revealed market and real market or payment) were crossed with validity terms (external validity, criterion validity or predictive validity). Appropriate mesh terms were used and the search strategy adapted for each of the databases (see Appendix 1 for a sample search strategy). In addition, reference lists of key papers and citation searches were conducted to identify additional papers. Results were handled using Mendeley reference management software.

### 2.2. Study selection criteria

The database search was run by one reviewer (LK) with reference lists and citation searches conducted by two reviewers (LK & JFR). All titles and abstracts, and full papers when in doubt, were double-reviewed (LK & JFR) using the following inclusion criteria: (1) conducted and reported in English; (2) assessed criterion validity of WTP/WTA; (3) included direct WTP elicitation methods (CV) only in both hypothetical and actual surveys; (4) included both a hypothetical and actual survey (with accompanying transaction) and (5) reported

empirical WTP or WTA values.

### 2.3. Data extraction

Data were extracted by one reviewer (LK) using a standard template in MS Excel (see Appendix 2), with a second reviewer double extracting data for a randomly selected 10% sample (SS). Disagreements were resolved through discussion, with any implications followed through to all other papers. Extracted data included background characteristics (e.g. country, terminology used, good valued), survey design (e.g. welfare perspective, elicitation format and pre-specified values for both hypothetical and actual WTP surveys where appropriate, payment vehicle, mode of administration, survey setting), study design (e.g. sampling (unit, sample selection, type of sample, size, response), duration between hypothetical and actual surveys, analytic methods (e.g. WTP estimation methods, regression methods) and main findings (types of comparisons produced and values). Where multiple comparisons were reported in a study, these were extracted separately. This was done to allow for the use of all the estimates and hence a larger dataset for analysis.

### 2.4. Risk of bias

A quality rating was not employed for individual studies as no agreed criteria exist for criterion validity assessments. Risk of bias, which could potentially affect the pooled results, was considered. Meta-bias (publication and selective reporting bias) was investigated using funnel plots (Lipsey and Wilson, 2001). The metafunnel command in Stata was used to explore the relationship between the ratio (logratio) and the standard error of the ratio (standard error of the logratio). In the absence of publication bias, the funnel plot generated from the studies included in the analysis should be inverted or asymmetrical. Where this is the case, the largest samples would be at the top of this inverted funnel plot, and closer to the true effect size. On the other hand, the smaller studies would be scattered along the x-axis. The reverse is true where publication bias is suspected.

### 2.5. Statistical analysis

For all comparisons, WTP estimates for hypothetical and actual data were matched as pairs, when provided, and compared as a ratio (for mean values) and as odds ratios (for percentage summaries). All quantitative analysis were conducted using Stata14 (StataCorp, 2015). Three types of analysis were conducted.

#### 2.5.1. Narrative summary

Using the entire dataset, a narrative and quantitative summary of the methods used in the comparisons and findings is provided. The comparisons of hypothetical and actual values in terms of background characteristics, survey design, study methods and results were summarized using counts, descriptive statistics, 2 by 2 tables, and box and whisker plots.

#### 2.5.2. Meta-analysis

A reduced dataset was used in the meta-analysis. For the mean summaries, only comparisons which reported standard errors of the mean, or those which provided sufficient statistics to enable the calculation of the standard error were included. Only comparisons which had a non-zero hypothetical and actual WTP value (and hence a non-zero odds ratio) were included in the meta-analysis for percentage summaries.

Given the variation in the methods used in the reviewed studies, a random-effects meta-analysis was conducted to calculate the weighted average of the log ratios and odds ratios (separately for mean and percentage summaries respectively). The weights were based on the inverse of variance of the effect estimates. Forest plots are presented separately for these and the $I^2$ statistic used to determine the level of heterogeneity (Higgins et al., 2003). Sensitivity analyses and subgroup analyses were also conducted in exploring the sources of the heterogeneity. In the sensitivity analysis, meta-analyses were re-run excluding comparisons with the smallest sample sizes. Sub-group analyses explored heterogeneity by sector, sample selection types, study administration modes and survey elicitation formats using the metan command in STATA14.

#### 2.5.3. Meta-regressions

Meta-regressions were conducted to explain the heterogeneity in the presented summaries and determine the drivers of hypothetical bias. These regressions were all clustered by study to control for the multiple comparisons from some of the studies

The dependent variables in these regressions were: (1) the ratio of hypothetical to actual values derived from comparisons presenting mean summaries, and (2) the log of the odds ratio of hypothetical to actual values, for comparisons presenting summaries as percentages. Previous meta-analyses investigated the effect of different study attributes on hypothetical bias (Carson et al., 1996; Harrison and Rutström, 2008; Liljas and Blumenschein, 2000; List and Gallet, 2001; Little and Berrens, 2003; Murphy et al., 2004). The results of these have been either mixed or inconclusive. In the absence of a theory explaining the divergence between hypothetical and actual WTP payments (hypothetical bias), the following variables were introduced into the models in an exploratory manner: (1) sector within which a valuation good or service falls; (2) class of good; (3) purpose of good; (4) study administration mode; (5) sample selection in both surveys; (6) type of sample (student or otherwise and users versus non-users of a service or good); (7) WTP elicitation format used in both surveys; (8) type of comparison (either between samples or within same sample); (9) study setting (laboratory or field); (10) duration between the hypothetical and actual surveys and (11) money effects (whether respondents were paid to participate in either survey or given money to purchase the good valued).

*2.5.3.1. Univariate regressions*. The range of univariate regressions explored the relationship between the dependent and independent variables listed in the previous section separately for comparisons presenting mean and percentage summaries. Significant variables are presented in the results section and discussed thereafter

*2.5.3.2. Multiple regression*. Where the ratio is the dependent variable (comparisons presenting mean summaries), the GLM estimator was used. The GLM permits the use of the estimates in their natural form, with a straight forward interpretation. Where the odds ratio was presented, the natural log was used and a logit model estimated. Base and reduced models were determined separately for comparisons summarized as means and percentages. In the base models, all the independent variables listed in section 2.4.3 were included. To arrive at a reduced model, variables with the highest non-significant p-values were removed and the model re-estimated. To examine model fit, model diagnostics were run with every estimation. The linktest (Cameron and Trivedi, 2009) was used to examine specification errors in the models. Further, the Hosmer Lemeshow test was used to check for the goodness of fit of the models (Hosmer and Lemeshow, 2013). The final reduced models included the range of variables which were significant and for which the models were best specified. Finally, for each of the models, a predicted ratio or log odds ratio was determined for the mean and percent summaries respectively

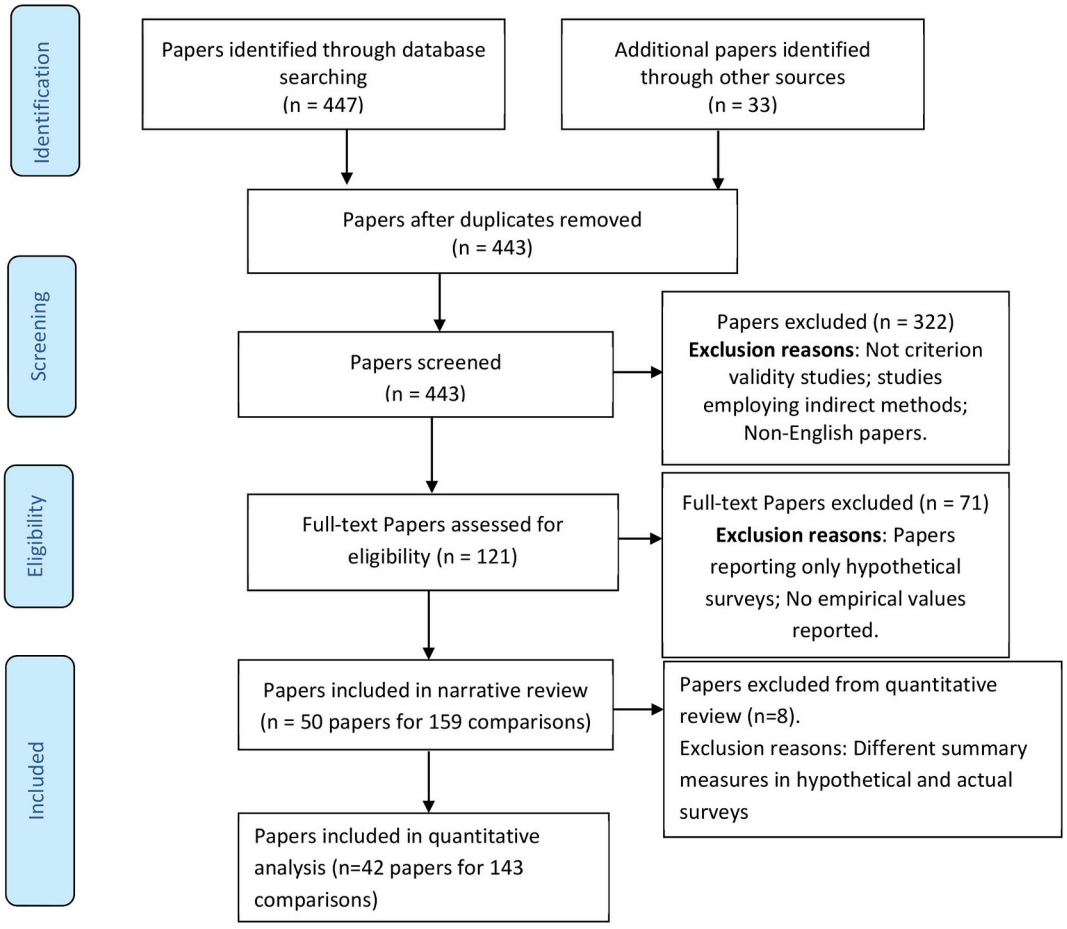



**Fig. 1.** Flow of papers during the search process.

## 3. Results

### 3.1. Background characteristics

Of the 480 papers identified, 50 were included (see Fig. 1) from 14 countries. Comparisons were typically carried out in the USA (n = 79 comparisons), followed by Norway (n = 35 comparisons), Nigeria (n = 16 comparisons) and Sweden (n = 9 comparisons). More than half the papers (n = 33) generated multiple comparisons (range: 2–30) of hypothetical and actual values. The results therefore, with the exception of country and year of publication, focus on 159 comparisons of hypothetical and actual WTP (WTA) values. Background characteristics of all the comparisons included in the review are provided in appendix 3.

The majority of comparisons (n = 94), did not explicitly use any specific terms for validity assessment, preferring to reflect papers as testing comparisons between hypothetical and actual WTP values. Approximately one fifth (n = 32) referred to this as testing for hypothetical bias (Blumenschein et al., 2014; Botelho and Pinto, 2002; Bryan and Jowett, 2010; Camacho-Cuena et al., 2004; Getzner, 2000; Johannesson, 1997; Mozumder and Berrens, 2007; Murphy Stevens et al., 2002; Onwujekwe et al., 2005). Two comparisons from the same study used the term predictive validity (Onwujekwe, 2001), while one used external validity (Muller and Ruffieux, 2011). A further one-fifth (n = 30) of the comparisons referred to assessments of criterion validity (Bhatia and Fox-Rushby, 2003; Bratt, 2010; Carlson, 2000; Johnston, 2006; Loomis et al., 1996; Onwujekwe et al., 2001; Onwujekwe and Uzochukwu, 2004; Onwujekwe, 2004; Ramke et al., 2009; Vossler et al., 2003a,b; Vossler and Kerkvliet, 2003; Willis and Powe, 1998).

**Table 1**
Type of good valued by sector.

| Type of good | Health | Environment | Other[a] | Total |
|---|---|---|---|---|
| Pure Public | 0 | 50 | 5 | **55** |
| Quasi-private | 0 | 10 | 14 | **24** |
| Pure Private | 36 | 0 | 44 | **80** |
| Total | **36** | **60** | **63** | **159** |

[a] Other sector includes goods or services that do not fall under the environment or health sectors.

Table 1 shows that most comparisons (38%) were in the environmental and 23% in the health sector, with the remainder spread in 'other' sectors. Of the 36 health sector comparisons, 30 elicited values for prevention products such as treated mosquito nets (Bhatia and Fox-Rushby, 2003) and six elicited values for management or treatment of a disease condition (e.g. Asthma management program (Blumenschein et al., 2001) and spectacles (Ramke et al., 2009)). In the environmental sector, 55 comparisons provided values for conservation, 2 elicited values for prevention purposes while 3 elicited values for use or access to public goods or services e.g. provision of public water to a remote village in Rhode Island (Johnston, 2006). Most comparisons (n = 54) in 'other' sectors elicited values for personal and household goods (e.g. art prints (Loomis et al., 1997; Loomis et al., 1996), sunglasses (Blumenschein et al., 2014)), one study elicited values for a personal good (chocolate bar) and a public good (prevention of additional damages to an aquatic system from acid rain) (Kealy et al., 1990).

**Table 2**
Comparison of study attributes in hypothetical and actual surveys.

| | Actual Survey | | | | | | |
|---|---|---|---|---|---|---|---|
| | Elicitation format | Auction | Bidding game | Dichotomous Choice | Open ended | Others[a] | Total |
| **Hypothetical Survey** | Auction | **11** | 0 | 0 | 1 | 0 | 12 |
| | Bidding game | 0 | **7** | 4 | 0 | 0 | 11 |
| | Dichotomous Choice | 0 | 0 | **66** | 2 | 2 | 70 |
| | Open ended | 12 | 0 | 4 | **27** | 0 | 43 |
| | Others[a] | 0 | 0 | 14 | 0 | **9** | 23 |
| | Total | **23** | **7** | **88** | **30** | **11** | **159** |
| | ***Survey administration mode*** | *In-Person* | *Mail* | *Self-administered* | *Telephone* | | *Total* |
| | In-Person | **96** | 3 | 0 | 0 | | **99** |
| | Mail | 5 | **47** | 0 | 2 | | **54** |
| | Self-administered | 2 | 0 | **0** | 0 | | **2** |
| | Telephone | 2 | 2 | 0 | **0** | | **4** |
| | Total | **105** | **52** | **0** | **5** | | **159** |
| | ***Sample selection*** | *Convenience* | *Purposive* | *Random* | | | *Total* |
| | Convenience | **48** | 6 | 0 | | | **54** |
| | Purposive | 2 | **66** | 10 | | | **78** |
| | Random | 0 | 6 | **21** | | | **27** |
| | Total | **50** | **78** | **31** | | | **159** |
| | ***Sample type*** | *Mixed* | *Non Students* | *Students* | | | *Total* |
| | Mixed | **2** | 0 | 0 | | | **2** |
| | Non Students | 0 | **116** | 1 | | | **117** |
| | Students | 0 | 0 | **40** | | | **40** |
| | Total | **2** | **116** | **41** | | | **159** |

[a] Other elicitation formats include all other elicitation formats with a count of less than 5 e.g. structured haggling, payment cards and mixed methods such as binary or bidding game with follow up.

### 3.2. Comparison of hypothetical and actual survey attributes

All comparisons adopted cross-sectional designs. Nearly all elicited WTP estimates (n = 154) while WTA values were derived in 5. One, in the environment sector (Heberlein and Bishop, 1986) sought WTA values in exchange for goose permits which hunters had earlier purchased in the hypothetical survey. In the actual survey, cash offers were made to the hunters to give up their permits. Four WTA comparisons were conducted in other sectors and these included eliciting expected compensation values from respondents in exchange for the holiday gifts followed by offers of actual payments for their holiday gifts (List and Shogren, 2002) and WTA in exchange for goose and deer permits (Heberlein and Bishop, 1986).

All comparisons used the same payment vehicle in actual and hypothetical surveys. Out of pocket payments were used in 154 comparisons across all sectors (exclusively so for the health and other sectors) and these included user fees and voluntary donations. Tax payments, primarily property taxes were used in 3 comparisons eliciting WTP values for public goods in the environmental sector (Vossler et al., 2003a,b; Vossler and Watson, 2013; Vossler and Kerkvliet, 2003). In the same sector, two comparisons were asked for voluntary donations towards a public good (Macmillan et al., 1999; Veisten and Navrud, 2006).

The majority of comparisons also used the same elicitation format (n = 111), administration mode (n = 143), sample selection technique (n = 135) and sample type (n = 158) in both the hypothetical and actual surveys. These are presented in Table 2 where, for every attribute, the diagonal in bold represents the similarities between hypothetical and actual surveys.

Different WTP elicitation formats were used across the hypothetical and actual surveys in nearly one-quarter of the comparisons (n = 39), where for example, the bidding game was used in the hypothetical survey but a dichotomous choice was used in the actual survey (Bhatia and Fox-Rushby, 2003; Vernazza et al., 2015). In one particularly unusual case, an open ended question is asked in the hypothetical survey, but an auction is used in the surveys of actual values (Fox et al., 1998). It is typically the environment (n = 54), and other (n = 43) sectors that have used the same elicitation formats for both the hypothetical and actual surveys. WTP for health goods/services has most commonly used different elicitation formats for the hypothetical and actual surveys (90%).

The same mode of administration, in-person interviews, was predominantly used in the "other" sectors but different modes of administration were used in the health and environment sector. For example, in the health sector, one study used mail surveys in the hypothetical survey but in-person interviews in the actual survey (Loomis et al., 2009). In the environment sector, four comparisons used mail surveys for hypothetical values and in-person interviews to elicit actual values (Vossler and Watson, 2013; Vossler and Kerkvliet, 2003; Johnston, 2006) with three comparisons (2 studies) using the opposite (Brown and Taylor, 2000; Seip and Strand. 1992).

Considering hypothetical and actual surveys separately, the general response rate was not indicated for nearly two-fifths of the comparisons in the hypothetical surveys (n = 63). However, in the actual survey, the general response rates were indicated in more than half the comparisons (n = 130). A comparison of the general response rates by study modes of administration shows that telephone interviews had the higher mean response rates, followed by mail surveys. Response rates from in-person interviews were scattered across the scale suggesting missing response values or outliers (Fig. 2).

The response rates to the valuation question were reported in only one-third of the comparisons for the hypothetical survey (n = 53) and for only fifteen comparisons in the actual survey. For thirteen comparisons (3 in health; 6 in environment; 4 in other sectors), the response rate for the actual and hypothetical questions was the same. Overall, the presence and treatment of the different non-responses, where

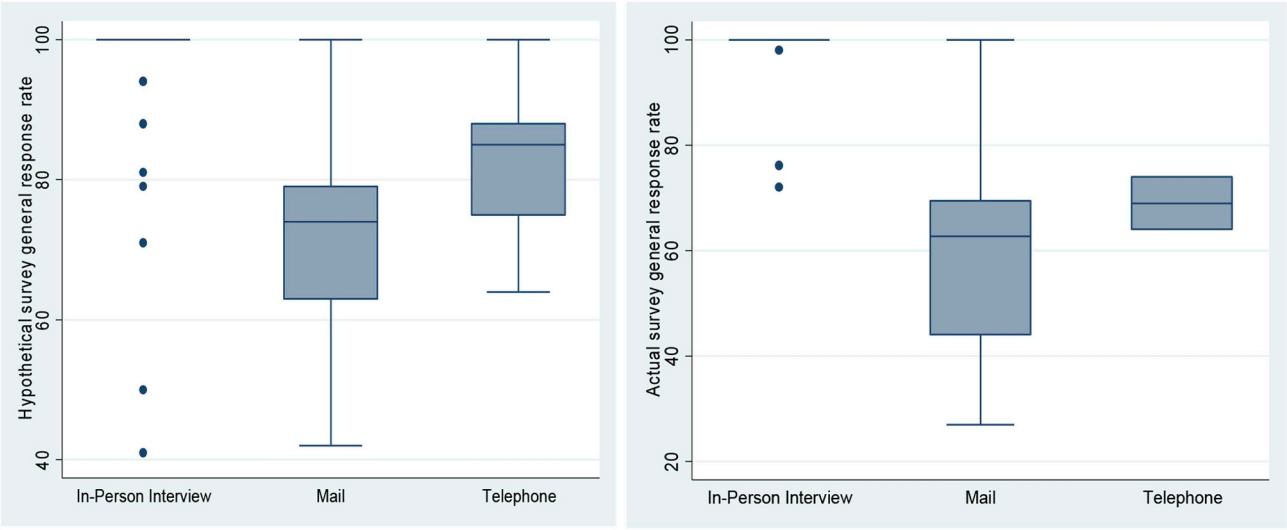



**Fig. 2.** Response rates by survey administration modes.

present, is not discussed. It is therefore not clear whether summary statistics provided exclude these missing values or not.

Fig. 3 compares the sample sizes used in the hypothetical and actual surveys, with five comparisons which were outliers dropped from the summary. Sample sizes ranged from 9 to 2890 in the hypothetical surveys and from 9 to 15,781 in the actual surveys. The sample sizes for the two surveys were similar in 88 comparisons. In most cases, where the sample size differed, the hypothetical survey had a larger sample than the subsequent survey of actual values (n = 44). However, for three comparisons from one study valuing a public good (comprehensive restoration plan for a riverfront commemorative park), the hypothetical survey sample size was less than 1% (122) of the actual survey sample size (15,781).

For more than two-fifths of the comparisons (n = 67) authors stated that different respondents were approached to complete the hypothetical and actual surveys, particularly so in the other (n = 35) and health sectors (n = 13). In most environment sector comparisons (n = 41/60) the same respondents were approached. Unfortunately, where the respondents and the sample size differ, tests relating to the representativeness of the sample of the actual survey in relation to the hypothetical survey were not always reported.

Hypothetical and actual surveys were undertaken at the same time or within a period of 2 weeks in the majority of comparisons (n = 126), with 31 administering the two surveys more than 2 weeks apart. The duration between the two surveys was not clear in 2 comparisons. The hypothetical and actual surveys conducted more than one month apart (n = 3) were in the environment sector (Vossler et al., 2003a,b; Johnston, 2006; Vossler and Watson, 2013).

### 3.3. Justification for the values used in the surveys

When closed-ended elicitation formats are used to elicit WTP or WTA values, pre-specified value cues are presented to respondents. For instance, a payment card presents a range of money values from which respondents are asked to select the value that best reflects their maximum WTP while bidding methods present single or multiple bids for valuation. As values presented are significant cues, they should not bias the true population mean WTP and therefore require justification to allow judgement of likely bias. However, in 56 comparisons across both the hypothetical and actual surveys for the same good, justifications were not provided for value cues used. In 7 comparisons from five studies, all in the environment sector, the values presented to the respondents in both the hypothetical and actual surveys were based on prior costings of the planned projects (Byrnes et al., 1999; Champ et al., 1997; Spencer et al., 1998; Champ and Bishop, 2001; Blumenschein Blomquist et al., 2008). In another sample, (Loomis et al., 1997), values obtained from a pre-test of the survey were presented to respondents in both surveys.

In four comparisons, values from hypothetical surveys were used to inform the actual survey (Bhatia and Fox-Rushby, 2003; Loomis et al., 1997; Willis and Powe, 1998; Onwujekwe, 2004). In two comparisons, one each in the health and environment sector, the stated hypothetical values were presented in the actual survey (Onwujekwe, 2004; Willis and Powe, 1998). One study each in the other sector (Loomis et al., 1997) and one in the health section (Bhatia and Fox-Rushby, 2003) used mean value from the hypothetical survey as the value cue for the actual surveys. Market prices for the commodities were used in the

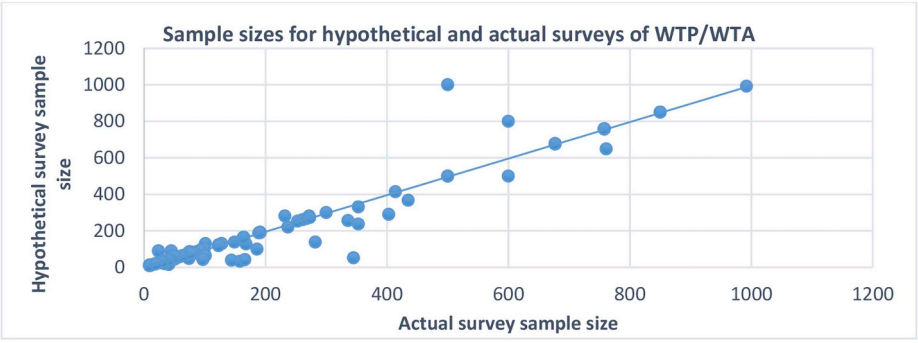



**Fig. 3.** Samples sizes of surveys for hypothetical and actual CV values.

actual surveys in one paper in the health sector (Onwujekwe et al., 2001). For 51 comparisons that used open ended questions, a justification was not relevant.

Most comparisons in the environment sector presented the stated hypothetical values in the actual survey (n = 34). In 11 comparisons, the value presented in the actual survey was based on a costing of the proposed project. In the 'other sectors', the market price for the good was presented in two comparisons while nearly one-third of the comparisons (n = 14) did not provide a justification for the values used in the hypothetical surveys. In two comparisons both from the same study (Loomis et al., 1997) the value presented in both hypothetical and actual surveys centred on a pre-test mean. Auctions and open ended elicitation formats were used in the actual surveys in 13 comparisons.

### 3.4. WTP/WTA estimates and criterion validity assessment

The estimation methods for mean WTP/WTA summaries are varied and these would be expected to relate to question format. Some studies, primarily those employing open ended questions, derived the summary estimates by a computation of averages (e.g., Balistreri et al., 2001; Fox et al., 1998; Getzner, 2000; List, 2001). The spread of this data is not given in around half (n = 50) of the comparisons presenting mean summaries. Summary WTP/WTA estimates were also modelled using a range of statistical techniques. Roughly 70% (n = 111) of the comparisons specified the statistical tests used with the majority (n = 88) employing parametric methods (non-parametric methods n = 18, both parametric and non-parametric methods n = 20). While similar elicitation formats were used in the hypothetical and actual surveys for nine comparisons, different summaries were presented. These included mean summaries in the hypothetical survey with a percentage in the actual survey and vice versa. 84 comparisons presented summary means for both surveys and 60 provided summary percentages. Different summary estimates were provided for 15 comparison pairs.

Study authors concluded that criterion validity was demonstrated where hypothetical and actual WTP estimates were relatively similar. However, the criteria for judging the ratios for a conclusion on criterion validity were often not provided. As a result, different conclusions were given even for similar ratios and odds ratios.

Criterion validity was not confirmed by study authors for more than three-quarters (n = 124) of comparisons. Of the 33 comparisons where study authors confirmed criterion validity of the WTP/WTA estimates, 17 were from the other sectors; 10 from the health sector and 6 from environment sector. Vernazza et al. (2015) reported mixed results for two comparisons in the health sector. Criterion validity confirmations were similar across the WTP summary methods. Table 3 summarises author's conclusions on criterion validity by sector and WTP/WTA summary measure.

Based on the summaries presented by the study authors, ratios (for mean comparisons) and odds ratios (for percentage summaries) were calculated. Of the comparisons that reported mean values in both the hypothetical and actual surveys (n = 84), the ratio of hypothetical to actual mean values was an average of 3.2 (range 0.7–11.8). The highest ratios were for environment sector (5.99), pure public (4.92), and conservation goods (5.96). For example, in one study which elicited WTP for the protection of sensitive rainforest land, the hypothetical mean WTP was $27.97 for female and $72.22 for male respondents whereas the mean actual WTP was $3.23 among females and $6.14 for males (Brown and Taylor, 2000). Ratios were also highest when the hypothetical and actual surveys were administered concurrently (3.24), when a donation mechanism was used as the payment vehicle (4.53) and when a one-off payment was elicited (3.22).

**Table 3**
Author-confirmations of criterion validity by sector and summary measure.

| | Criterion Validity Confirmations (total no. of summaries) | | | |
|---|---|---|---|---|
| Sector | Mean Summaries | Percentage Summaries | Mixed Summaries | Total |
| Environment | 2 (23) | 2 (33) | 2 (4) | **6** |
| Health | 0 (9) | 9 (16) | 1 (11) | **10** |
| Other | 15 (52) | 2 (11) | | **17** |
| **Total** | **17 (84)** | **13 (60)** | **3 (15)** | **33 (159)** |

**Table 4**
Summary estimates by study attribute (Overall).

| Variable | Ratio [SD] *(no. of comparisons)* | Odds ratio [SD] (no. of comparisons) |
|---|---|---|
| Sector | | |
| i. Health | 1.75 [**0.70**] *(9)* | 0.29 [**0.43**] *(15)* |
| ii. Environment | 5.99 [**3.75**] *(23)* | 0.88 [**2.04**] *(30)* |
| iii. Other[a] | 2.27 [**3.47**] *(52)* | 1.30 [**0.81**] *(11)* |
| Class of good or service | | |
| i. Pure Public | 4.92 [**3.68**] *(22)* | 1.24 [**0.86**] *(26)* |
| ii. Pure Private | 2.49 [**3.67**] *(42)* | 0.72 [**0.79**] *(26)* |
| iii. Quasi-Private | 2.70 [**3.31**] *(20)* | −1.48 [**5.03**] *(4)* |
| Purpose of good or service | | |
| i. Prevention | 1.57 [**0.45**] *(8)* | 0.13 [**0.18**] *(13)* |
| ii. Conservation | 5.96 [**3.78**] *(23)* | 0.76 [**1.97**] *(28)* |
| iii. Treatment | 3.25 [-] *(1)* | 0.81 [**0.44**] *(6)* |
| iv. Other[b] | 2.19 [**3.34**] *(52)* | 1.92 [**1.25**] *(9)* |
| Type of comparison | | |
| i. Between | 3.12 [**4.06**] *(53)* | −0.32 [**2.67**] *(11)* |
| ii. Within | 3.27 [**3.04**] *(31)* | 1.08 [**1.03**] *(45)* |
| Survey setting | | |
| i. Field | 0.98 [**0.83**] *(29)* | 0.69 [**1.61**] *(50)* |
| ii. Laboratory | 0.66 [**0.75**] *(55)* | 1.72 [**0.81**] *(6)* |
| Duration between surveys | | |
| i. Concurrent | 3.24 [**3.86**] *(76)* | 1.32 [**0.97**] *(25)* |
| ii. 1–7 days | 2.78 [**1.34**] *(7)* | 1.54 [**0.98**] *(12)* |
| iii. More than 7 days | 1.08 [-] *(1)* | −0.33 [**1.91**] *(19)* |
| Payment Vehicle | | |
| i. Cash Fee | 2.83 [**3.69**] *(67)* | 0.89 [**1.07**] *(29)* |
| ii. Donation | 4.53 [**3.51**] *(17)* | 1.27 [**0.85**] *(24)* |
| iii. Property tax | | −3.72 [**3.23**] *(3)* |
| Payment duration | | |
| i. Annual Payment | 1.39 [**0.44**] *(2)* | 0.95 [**1.33**] *(2)* |
| ii. One-Off | 3.22 [**3.73**] *(82)* | 1.04 [**0.99**] *(52)* |
| iii. Monthly | – | −5.58 [**0.10**] *(2)* |
| Overall | 3.17 [**3.70**] *(84)* | 5.72 [**1.57**] *(56)* |

[a] Other sectors includes consumer goods such as books, sunglasses.
[b] Includes consumables such as food, clothing and household items.

For the comparisons which presented percentage summaries in both hypothetical and actual surveys odds ratios were calculated for only the comparisons which had non-zero values in both surveys (n = 56). The average odds ratio was 5.7 (range of 0–13.6). The highest odds ratios were observed in; comparisons in the environment sector (0.88), quasi private goods (−1.48), goods used for "other" purposes (1.96), within sample comparisons (1.08), studies conducted within a laboratory setting (1.72), study periods of between 1 and 7 days between the hypothetical and actual surveys, when a property tax was used as the payment vehicle (−3.72) and when monthly payments were elicited (−5.58). The ratios and odds ratios for the included comparisons by different design attributes are presented in Table 4 (overall characteristics) and Table 5 (hypothetical and actual surveys).

In the comparisons of hypothetical and actual surveys (Table 5), the highest ratios were observed when a purposive sample was used in the hypothetical survey (3.60); a random sample in the actual survey (4.22); with a mixed (student and non-student) sample in both the hypothetical and actual surveys (10.21 in both), followed by a non-student sample in both surveys; with the use of telephone surveys in the hypothetical survey (5.60) and open ended surveys in the hypothetical (5.10) and actual (5.49) surveys. The sample type (whether users or non-users of the valuation good/service) generated similar ratios in both the hypothetical and actual surveys.

The highest odds ratios were observed for percentage summaries where; convenience samples were used in both the hypothetical and

**Table 5**
Summary estimates by study attribute (hypothetical and actual surveys).

| Survey Attributes | Ratio [SD] *(no. of observations)* | | Odds ratio (no. of observations) | |
|---|---|---|---|---|
| | Hypothetical Survey | Actual Survey | Hypothetical Survey | Actual Survey |
| Sample Selection | | | | |
| i. Random | 1.27 [**0.37**] *(4)* | 4.22 [**3.37**] *(11)* | −0.19 [**1.97**] *(19)* | −0.27 [**2.22**] *(15)* |
| ii. Purposive | 3.60 [**3.25**] *(32)* | 2.86 [**2.73**] *(31)* | 1.24 [**1.05**] *(31)* | 1.11 [**1.05**] *(35)* |
| iii. Convenience | 3.05 [**4.10**] *(48)* | 3.13 [**4.37**] *(42)* | 1.72 [**0.81**] *(6)* | 1.72 [**0.81**] *(6)* |
| Sample type (1) | | | | |
| i. Students | 2.98 [**4.54**] *(33)* | 2.94 [**4.48**] *(34)* | 1.72 [**0.81**] *(6)* | 1.72 [**0.81**] *(6)* |
| ii. Non-Students | 3.02 [**2.77**] *(49)* | 3.05 [**2.80**] *(48)* | 0.69 [**1.61**] *(50)* | 0.69 [**1.61**] *(50)* |
| iii. Mixed | 10.21 [**2.19**] *(2)* | 10.21 [**2.19**] *(2)* | – | – |
| Sample type (2) | | | | |
| i. Users | 2.99 [**3.66**] *(75)* | 2.99 [**3.66**] *(75)* | 0.83 [**1.59**] *(54)* | 0.80 [**1.57**] *(56)* |
| ii. Non-Users | 4.75 [**3.84**] *(9)* | 4.75 [**3.84**] *(9)* | −0.02 [**0.05**] *(2)* | – |
| Study administration mode | | | | |
| i. In-person interviews | 2.78 [**3.71**] *(60)* | 2.47 [**3.40**] *(61)* | 0.89 [**1.07**] *(29)* | 0.82 [**1.06**] *(29)* |
| ii. Mail Surveys | 3.88 [**3.38**] *(20)* | 5.05 [**3.87**] *(23)* | 0.71 [**2.00**] *(27)* | 1.29 [**0.84** *(25)* |
| iii. Telephone | 5.60 [**4.66**] *(4)* | – | – | −5.58 [**0.10**] *(2)* |
| Survey elicitation format | | | | |
| i. Auction | 1.74 [**0.83**] *(12)* | 3.05 [**4.88**] *(23)* | – | – |
| ii. Bidding game | 1.60 [-] *(1)* | 1.60 [-] *(1)* | 0.11 [**0.14**] *(6)* | 0.09 [**0.15**] *(5)* |
| iii. Dichotomous choice | 3.08 [**2.54**] *(29)* | 2.18 [**1.72**] *(34)* | 0.67 [**1.68**] *(37)* | 0.73 [**1.72**] *(41)* |
| iv. Open ended | 5.10 [**5.40**] *(27)* | 5.49 [**4.14**] *(17)* | 1.90 [**1.41**] *(10)* | 1.59 [**1.07**] *(9)* |
| v. Payment Card | 1.17 [**0.42**] *(7)* | 3.05 [**3.75**] *(9)* | – | – |
| vi. Referendum | 1.15 [**0.24**] *(8)* | – | – | – |
| vii. Bidding + OE | – | – | −0.60 [-] *(1)* | – |
| viii. Binary with follow up | – | – | 0.25 [**0.11**] *(2)* | 0.16 [-] *(1)* |

actual surveys (1.72 in each), with students were sampled (1.72 in each); respondents were potential users of the valuation good (0.83 in the hypothetical and 0.80 in the actual surveys); telephone interviews were used in the actual survey and with open ended survey elicitation formats (1.59).

### 3.5. Results of meta-analyses

Meta-analysis was conducted separately for comparisons presenting mean and percentage summaries. Standard errors were provided or calculated where possible for a total of fifty four of the comparisons presenting mean summaries and only these were included in the meta-analysis. For comparisons presenting percentage summaries, four reported a zero value in the actual survey results, generating an odds ratio of zero. These were excluded from the analysis, leaving a total of 56 comparisons. Fifteen comparisons which presented different summaries in the hypothetical and actual surveys were also excluded from the meta-analysis.

#### 3.5.1. Comparisons presenting mean summaries

The ratio of the actual and hypothetical mean values was used in the random effects meta-analysis. The pooled ratio of hypothetical to actual WTP values for the 54 comparisons was 1.79 with a range of 1.56–2.04 (see Fig. 4). This implies that for these comparisons hypothetical WTP was higher than actual WTP by 79%. Some variation in the effect sizes was expected, given the differences in the characteristics of the comparisons pooled in this analysis. However, a very high level of heterogeneity was detected in pooling the 54 comparisons. This is indicated by the $I^2$ of 97.1% which was significant ($p < 0.001$). .

The pooled ratio (see Fig. 4) was highest in the environment sector (1.85) compared to 1.25 in the other and 1.49 in the health sectors (Appendix 4). In addition, studies in the health sector had the lowest heterogeneity level (56.5%, $p = 0.0056$). However, the number of comparisons in the health sector was small (4) and from the same study. This compares with heterogeneity levels in the environment (92.7%, $p < 0.001$) and other (97.7%, $p < 0.001$) sectors.

In the subgroup analysis by survey setting, while the overall level of heterogeneity remained high and significant regardless of study setting (97.1%, $p < 0.001$ overall), this was much lower with field studies (68.4%, $p < 0.001$) compared to laboratory studies (97.7%, $p < 0.001$) (Appendix 5). In a sensitivity analyses, the effect on the pooled ratio of dropping comparisons which had the widest confidence intervals was explored. The pooled ratio was slightly smaller at 1.78 but the level of heterogeneity increased by 0.3 percentage points, remaining significant (97.4%, $p < 0.001$) (Appendix 6).

#### 3.5.2. Comparisons presenting percent summaries

The log odds ratio of the actual and hypothetical percentages was used in the random effects meta-analysis. A forest plot of these comparisons is presented in Fig. 5. The forest plot shows that respondents were more likely to say "yes" in the hypothetical survey than they were in the actual survey. The pooled odds ratio from the studies presenting percent summaries was 2.37 (range 1.93–2.80) i.e. the odds of saying "yes" in the hypothetical survey were more than double the odds of saying "yes" in the actual survey. As the level of heterogeneity was high and significant (90.2%, $p < 0.001$), the variation could not be attributed to chance alone.

Sub-group analysis showed heterogeneity was high and significant for studies from the environment sector (93.25%, $p < 0.001$). Heterogeneity in the other and health sectors was considerably lower and insignificant (35.2%, $p = 0.117$ & 21.9%, $p = 0.211$ respectively), and that the variation could be attributed to chance alone (Appendix 7). The differences in the levels of heterogeneity were not significant for the other study attributes. The differences by survey setting (Appendix 8) could be attributed to the few laboratory studies.

In the sensitivity analysis, three comparisons which had the widest confidence intervals were dropped from the analysis. In the meta-analysis of this reduced sample, the pooled odds ratio from the comparisons was slightly higher (2.36) and significant ($p < 0.001$) (Appendix 9).

The pooled estimates from both the mean and percentage summaries demonstrated that hypothetical WTP estimates overestimate actual values. For all the analyses presented, high levels of heterogeneity were noted. The explorations of the heterogeneity did not isolate any study characteristic as contributing to this. While this suggests that the variation in the estimates across the comparisons was not due to chance, this might simply be due to the differences in the pooled studies. Using

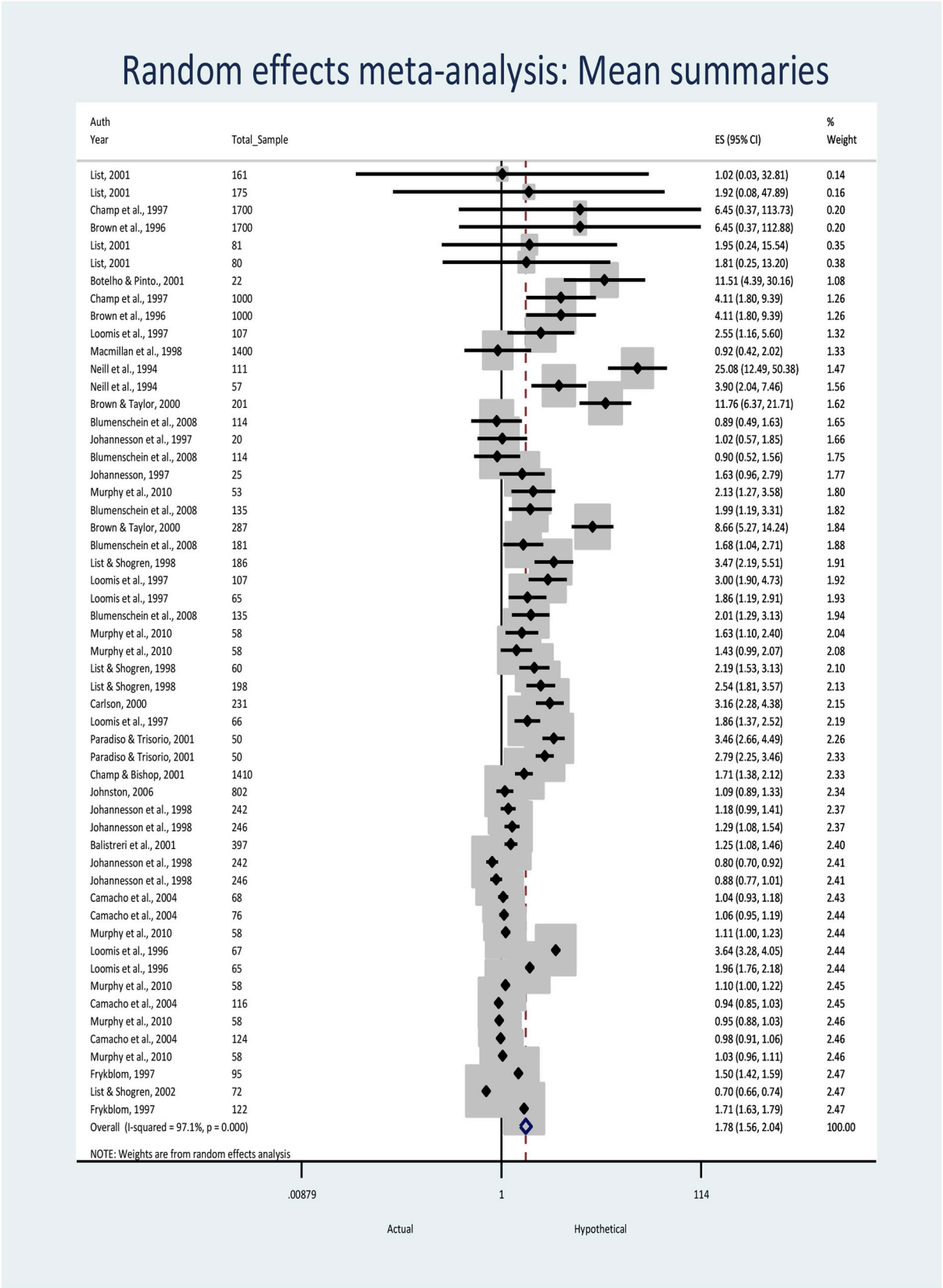



**Fig. 4.** Random effects meta-analysis_Mean summaries.

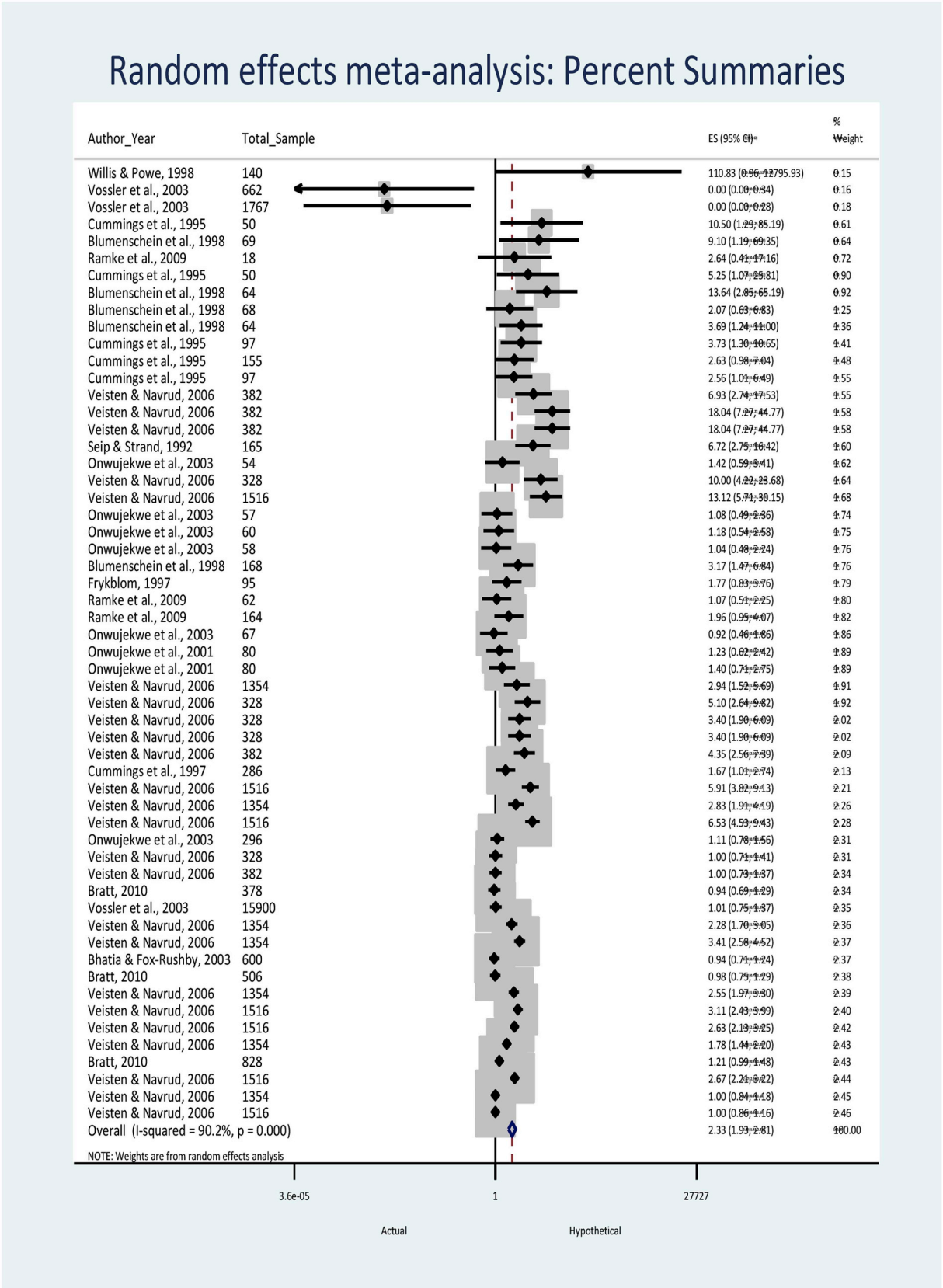



**Fig. 5.** Percent summaries meta-analysis.

**Table 6**
Univariate meta-regression outputs for mean and percent summaries.

| Variables (reference category) | Coeff (s.e.) Mean: Ratio | Coeff (s.e.) Percent: Log Odds Ratio |
|---|---|---|
| *Study sector (No)* | | |
| Other | −2.483***(0.797) | 0.615*(0.350) |
| Health | **−1.706***(0.501)** | **−0.701**(0.303)** |
| Environment | 3.878***(0.866) | 0.162 (0.405) |
| *Type of good (No)* | | |
| Pure Public | **2.363***(0.892)** | **0.822**(0.391)** |
| Pure Private | −1.373*(0.794) | −0.162(0.405) |
| Quasi Private | −0.626(0.870) | −2.475(2.223) |
| *Purpose of good (No)* | | |
| Conservation | 3.837***(0.873) | −0.0957(0.427) |
| Other | −2.590***(0.802) | 1.331***(0.464) |
| Prevention | **−1.777***(0.467)** | **−0.694**(0.326)** |
| Treatment | 0.0801(0.409) | 0.00767(0.291) |
| *Similarities in hypothetical and actual survey attributes (No)* | | |
| Sample type and characteristics in both surveys | | |
| Same sample type | −1.362 (0.895) | 0.737***(0.242) |
| Student sample | 0.294 (0.873) | 1.029**(0.387) |
| Random sample | **−2.005*** (0.453)** | **−1.482**(0.589)** |
| Purposive sample | −0.315 (0.790) | 0.976**(0.430) |
| Convenient sample | −0.0855 (0.808) | 1.029**(0.387) |
| Good or service users | −1.766 (1.287) | 0.864***(0.222) |
| *Survey administration in both surveys (No)* | | |
| Same administration mode | −3.378**(1.685) | 3.364*(1.707) |
| Mail mode | 1.104(0.898) | 0.813**(0.376) |
| In-person interview | −2.111**(0.869) | 0.0957(0.427) |
| *WTP elicitation format in both surveys (No)* | | |
| Same WTP elicitation format | −0.287(1.045) | −0.527(1.028) |
| Auction | −0.169(1.083) | – |
| Bidding | **−1.595***(0.409)** | **−0.782***(0.241)** |
| Dichotomous Choice | −0.920(0.664) | −0.447(0.418) |
| Open ended | **3.549***(1.171)** | **0.936**(0.415)** |
| Payment Card | −2.188***(0.460) | – |
| *Other survey attributes* | | |
| Duration between both surveys (non-concurrent) | 0.667 (0.638) | 0.930**(0.387) |
| One-off payment (*No*) | **1.823***(0.469)** | **3.364*(1.707)** |
| Payment vehicle cash fee (donation) | −1.697*(0.946) | 0.174(0.436) |
| Comparison type Between (Within) | −0.145(0.777) | −1.411*(0.800) |
| Survey setting Field (Laboratory) | 0.917(0.808) | −1.029**(0.387) |
| Money given for participation in either survey (*No*) | −0.607 (0.796) | 0.333(0.388) |
| Validity conclusion Confirmed (Not confirmed) | −2.556***(0.492) | −0.667**(0.316) |
| Observations | 84 | 56 |

Robust standard errors in parentheses ***p < 0.01, **p < 0.05, *p < 0.1.

meta-regressions, the drivers of this variation in the ratio and odds ratios were further explored. This analysis is presented in the next section.

### 3.6. Meta-regression results

All the comparisons presenting mean summaries are included in the regression analysis (n = 84) while only 56 comparisons presenting percentage summaries are included. Univariate and multiple regression results are presented separately in the next section. For the presented multiple regression models, the linktest estimate was not significant, indicating that the models were correctly fitted. In interpreting all the regression results, variables with positive coefficients are associated with higher ratios (odds ratios) and therefore higher hypothetical bias. Similarly, negative coefficients are associated with lower ratios (odds ratios) and therefore lower hypothetical bias.

**Table 7**
Meta-regression output for mean summaries.

| Variables (reference category) | Base model Ratio Coeff. (s.e) | Reduced model Ratio Coeff. (s.e) |
|---|---|---|
| *Sector (Health)* | | |
| Environment | 5.778***(1.590) | **4.290***(0.821)** |
| Other | 1.491(1.417) | 0.0659(0.490) |
| *Classification of valuation good (Pure private)* | | |
| Pure Public | 2.314(1.679) | **1.839**(0.757)** |
| Quasi Private | 1.910**(0.877) | **2.250***(0.767)** |
| *Purpose of good (Other)* | | |
| Conservation | 1.855,872 (0.179) | |
| Similarities in hypothetical and actual survey attributes (*No*) | | |
| Sample type and characteristics in both surveys | | |
| Same sample type | 0.768(1.033) | |
| Student sample | 0.314(1.187) | |
| Random sample | −2.457*(1.436) | **−3.401***(0.981)** |
| Purposive sample | 0.845(1.134) | |
| Convenient sample | – | |
| Good or service users | 2.462**(1.256) | **2.104**(0.971)** |
| *Survey administration in both surveys (No)* | | |
| Same administration mode | −2.677***(0.915) | **−2.601***(0.764)** |
| Mail mode | −3.187**(1.499) | **−1.795**(0.712)** |
| In-person interview | – | |
| *WTP elicitation format in both surveys (No)* | | |
| Same WTP elicitation format | −4.349(2.689) | **−3.802*(2.240)** |
| Auction | 5.586*(2.943) | **5.611**(2.349)** |
| Bidding | 0.820(3.110) | |
| Dichotomous Choice | 5.547*(3.064) | **4.936**(2.169)** |
| Open ended | 7.130**(3.071) | **6.428***(2.137)** |
| Payment Card | 4.892(3.168) | **4.586*(2.395)** |
| *Other survey attributes* | | |
| Duration between both surveys (non-concurrent) | 3.743**(1.775) | **3.308**(1.633)** |
| One-off payment (*No*) | 3.840***(1.403) | **4.500***(0.860)** |
| Payment vehicle cash fee (donation) | 0.612(1.076) | |
| Comparison type Between (Within) | −0.125(0.180) | |
| Survey setting Field (Laboratory) | 0.316*(0.185) | |
| Money given for participation in either survey (*No*) | −0.0308(0.457) | |
| Validity conclusion Confirmed (Not confirmed) | −2.075**(0.949) | **−2.130***(0.776)** |
| Link test (hatsq) | 0054 (1.43) | 0.059 (0.121) |
| Predicted mean ratio (s.d) | 3.179 (2.640) | 3.179 (2.640) |
| Observations | 84 | 84 |

Robust standard errors in parentheses ***p < 0.01, **p < 0.05, *p < 0.1.

#### 3.6.1. Univariate regression: Mean and Percent summaries

The sector within which the valuation good falls, the type and purpose of good were all significant with the direction of influence similar for both the mean and percent summaries. The ratio and log odds ratio were reduced for a good in the health sector, prevention goods and goods classified as pure private good. Both the ratio and log odds ratio for pure public goods were significantly higher.

In comparing similarities in design attributes for the hypothetical and actual surveys, both the ratio and log odds ratio were lower when the same WTP elicitation method was used in both surveys and with the use of the bidding technique. Random sampling techniques also contributed to lower ratios and log odds ratios. However, the use of open-ended WTP elicitation and one-off payments elicited significantly higher ratios and log odds ratios.

The direction of effect was reversed for mean and percent comparisons for some of the attributes. For instance, while a good in the other sectors elicited lower ratios, the log odds ratio was higher and both were significant. This could be explained by the number of comparisons in the mean and percent summaries. Similar effects and direction were seen for comparisons using the same sample type, administration mode and in-person interviews in both surveys. The univariate regression outputs for both mean and percent summaries are

**Table 8**
Meta-regression output for percent summaries.

| Variables (reference category) | Base model Log O.R Coeff. (s.e) | Reduced model O.R Coeff. (s.e) |
|---|---|---|
| *Sector (Health)* | | |
| Environment | −1.415*(0.709) | 0.379***(0.303) |
| Other | 0.358(0.452) | 1.271 (0.284) |
| *Classification of valuation good (Pure private)* | | |
| Pure Public | 15.65***(1.813) | ~213***(0.614) |
| Quasi Private | – | |
| *Purpose of good (Prevention)* | | |
| Conservation | −20.64***(2.939) | 0.0002***(1.020) |
| Other | −1.558**(0.734) | 1.001 (0.243) |
| *Similarities in hypothetical and actual survey attributes (No)* | | |
| Sample type and characteristics in both surveys | | |
| Same sample type | 8.504***(1.147) | 248.14***(0.403) |
| Student sample | – | |
| Random sample | −0.0907(0.501) | |
| Purposive sample | – | |
| Convenient sample | – | |
| Good or service users | −0.473(0.371) | |
| *Survey administration in both surveys (No)* | | |
| Same administration mode | −8.078***(1.477) | 0.0044***(0.639) |
| Mail mode | – | |
| In-person interview | – | |
| *WTP elicitation format in both surveys (No)* | | |
| Same WTP elicitation format | −7.230***(1.309) | 0.004*** (0.360) |
| Auction | – | |
| Bidding | −0.553(0.449) | |
| Dichotomous Choice | −0.481(0.442) | |
| Open ended | – | |
| Payment Card | – | |
| *Other survey attributes* | | |
| Duration between both surveys (non-concurrent) | −0.353(0.348) | |
| One-off payment (*No*) | | |
| Payment vehicle cash fee (donation) | −5.306***(1.559) | 0.026***(0.403) |
| Comparison type Between (Within) | 0.806**(0.366) | |
| Survey setting Field (Laboratory) | −0.995(0.855) | |
| Money given for participation in either survey (*No*) | 2.046***(0.374) | 3.161***(0.266) |
| Validity conclusion Confirmed (Not confirmed) | −0.0652(0.474) | |
| Link test (hatsq) | -.0005 (0.971) | 0.0001 (1) |
| Predicted odds ratio (s.d) | 0.808 (1.472) | 2.24 (1.44) |
| Observations | 56 | 56 |

Robust standard errors in parentheses ***p < 0.01, **p < 0.05, *p < 0.13.7.

presented in Table 6.

### 3.6.2. Meta-regression

The meta-regression results are presented separately for comparisons presenting mean and percent summaries (Tables 7 and 8). For both the base and reduced models presented, the regressions weighted by the study fit the data for $r^2$ of 0.68 and 0.65 respectively. Interestingly, the direction of effect is maintained in the base and reduced models for the mean summaries regression (with the exception of one variable), whereas there were five changes in sign for the percent summaries regression. The discussion focusses on the reduced model results as they fit the data better.

*3.6.2.1. Mean summaries*. The ratio of hypothetical to actual WTP values significantly increased for; goods in the environment sector compared to the health sector (4.3), pure public goods (1.8) and quasi-private goods (2.3) when compared to pure private goods, and valuation with good or service users (2.1) compared to non-users. Focusing on the WTP elicitation format, while the use of the same format in both the hypothetical and actual surveys reduced the ratio by 3.8 ($p < 0.1$), the use of open ended methods increased the ratio by 6.4 ($p < 0.01$) while auctions and bidding methods increased the ration by 5.6 ($p < 0.05$) and 4.9 ($p < 0.05$) respectively.

Conversely, the use of random sampling techniques significantly reduced the ratio of hypothetical to actual WTP values by a factor of 3.4 ($p < 0.001$). The use of the same administration mode in both surveys, and in particular, the use of mail surveys, significantly reduced the ratio by 2.6 ($p < 0.001$) and 1.7 ($p < 0.05$) respectively. As would be expected, the ratio of hypothetical to actual values was significantly lower where authors had concluded that criterion validity had been established based on their study findings (2.1, $p < 0.001$). Finally, based on the model, the predicted ratio for comparisons presenting mean summaries was 3.1 (s.d. 2.64). The model estimation results are presented in Table 7.

*3.6.2.2. Percent summaries*. For comparisons presenting percent summaries, the log odds ratios were back transformed into odds ratios for ease of interpretation (see Table 8). The odds of a higher WTP value in the hypothetical survey (and hence higher odds ratio) were statistically significantly higher for valuation goods classified as pure public, where the same sample was approached in both the hypothetical and actual surveys and when participants were given money to participate in either the hypothetical or actual survey (money effects). However, the odds of a higher hypothetical WTP value (lower odds ratio) were 62% lower when a valuation good was from the environment sector (compared to the health sector); more than 99% lower when the same administration mode and elicitation format were used in the hypothetical and actual surveys of WTP, and when cash was asked, compared to donations. All these results were statistically significant at $p < 0.001$. The predicted odds ratio for these comparisons was 2.24.

### 3.6.3. Summary of regression results: comparisons between mean and percent summaries

In comparing the characteristics of both surveys, using the same administration mode and WTP elicitation formats in the hypothetical and actual surveys led to lower ratios and odds ratios. The ratios and odds ratios for valuation goods classified as pure public and those in the other sectors were significantly high. Differences were observed for all the other variables in the reduced models.

### 3.7. Risk of bias analysis

Meta-bias was investigated separately for the studies reporting mean and percentage summaries. As illustrated in Fig. 6 (for studies reporting mean summaries) and Fig. 7 (for studies reporting percentage summaries), the funnel plots would signify the presence of publication bias, provided there was no difference in methods between the large and smaller studies. However, these funnel plots do not account for the differences in the methods used in the different studies. If the large studies differ in methods or other key characteristics, this relationship is not expected to hold. There are substantive differences in the methods used in the different studies. These differences are shown in this paper to affect the difference in the gap between the hypothetical and actual WTP. Therefore, these analysis of publication bias are very likely themselves biased.

## 4. Discussion

This review shows that considerable research has focussed on the criterion validity of CV methods since the late 1990s, with most papers from the health sector appearing after 2000. It is the first review and meta-analysis on criterion validity for over a decade and presents the first meta-regression that explores potential reasons for differences between hypothetical and actual WTP across studies. . With the increasing use of simulated market experiments, it is not surprising that

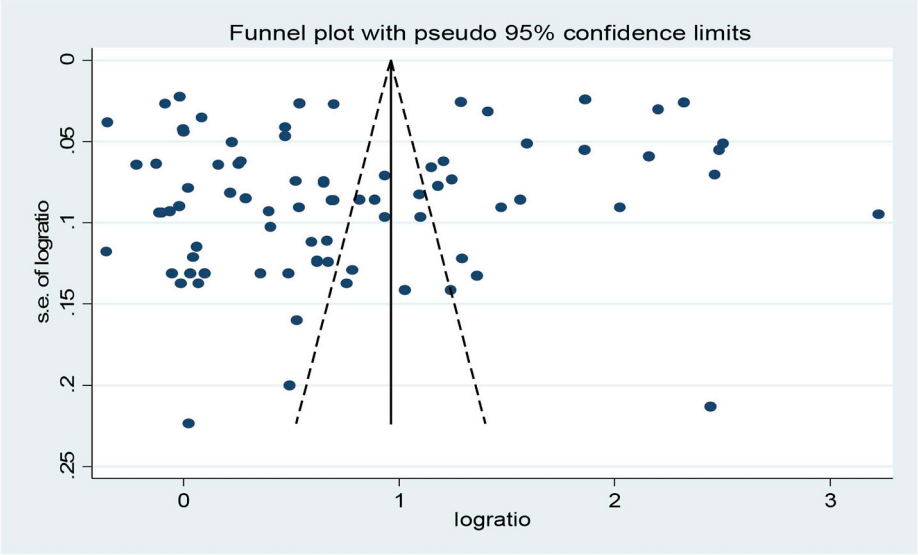



**Fig. 6.** Funnel plot of the ratio and standard errors of the log ratios (for mean summaries).

the majority of the work has focussed on private goods and this is particularly the case beyond the health and environment sectors. However, an important body of evidence now also exists for quasi-public and public goods/services. Applications in the environmental sector lead all assessments of criterion validity for public goods and the greater part of quasi-public goods. Almost two-thirds of investigations are for private goods in the US, with the remaining 35% spread across 9 countries. The question of whether results from simulated market experiments for a private good can transfer to evidence of the validity of CV methods in quasi- or pure-public goods has not yet been addressed.

The definition of external/criterion validity differs in the CV literature, but authors have equated this type of research with assessments of construct validity and reliability. This variety could explain why a large proportion of our evidence was accessed through reviews of references and citation searching. Previous reviews encountered the same difficulty and criterion validity assessments published since 2005 continue to use a variety of terms to describe similar types of research. Future reviews might therefore consider a wider variety of search terms but expect this to be resource intensive in the very large numbers of titles and abstracts returned for review.

This paper gives an indication of the degree of variation in hypothetical and actual WTP in the CV literature; hypothetical WTP (WTA) was on average 5.1 times greater (lower) than actual WTP (WTA), with a range of 1.5–11.99. The meta-analysed results place the degree of variation as 1.79 (range: 1.56–2.04) for mean summaries and 2.37 (range: 1.93–2.80) for percent summaries. Further, the predicted ratios and odds ratios of 3.18 and 2.24 respectively from the meta-regression further confirm the variation in stated and actual WTP values. The review also shows that current conclusions are heavily weighted (76% agreement) towards claims that criterion validity is not demonstrated, as only 24% authors claim evidence of criterion validity. Analysis for publication bias can, assuming no difference in methods between large and small studies, be used to question pooled evidence on the presence of criterion validity as it would indicate studies which demonstrated a lack of criterion validity were published whereas those that show validity are not. Whether this hegemony exists is tempered by our finding of great variety in methods, and it would be worth exploring whether a difference in methods between large and small studies would allay concerns over publication bias. However, alongside this evidence, we have found neither discussion of 'how close is close enough?' nor consideration of how valid the evidence itself was and therefore we question whether the results are quite as robust as they

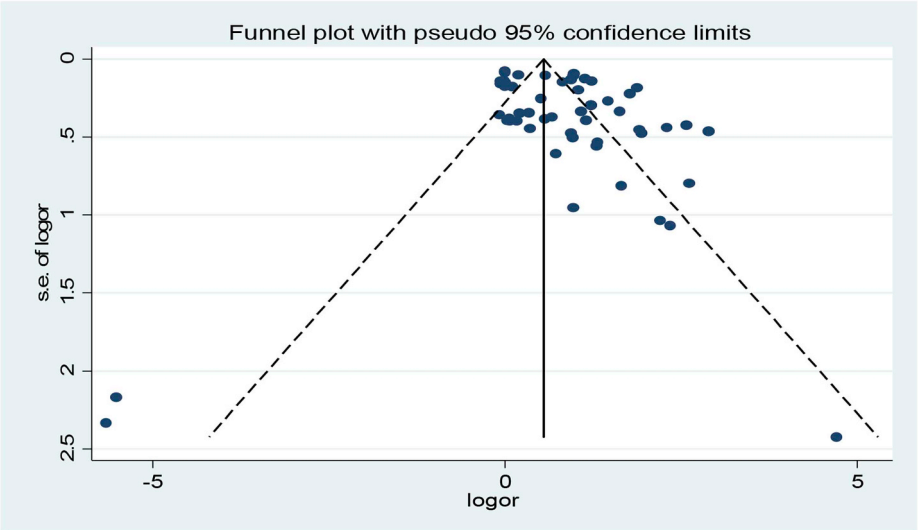



**Fig. 7.** Funnel plot of the odds ratio and standard errors of the log odds ratios (for percentage summaries).

appear to be.

This review has highlighted a great deal of methodological variation between hypothetical and actual surveys, and potentially sufficient variation to question the validity of findings about criterion validity itself. For example, the elicitation format was different in over half the comparisons, the same value cues were not necessarily used as results from some hypothetical surveys influenced values presented in the actual survey. A series of other differences relate to variation in the survey comparisons used between hypothetical and actual surveys. For example, half the papers stated that different populations were used and 54% clearly used different sample sizes. As all these differences have been shown to influence mean WTP (Trapero-Bertran et al., 2013; Veronesi et al., 2011), there could arguably be a good reason to accept that WTP results should infact be different. Comparisons also involved a wide range of goods and differing conclusions on criterion validity were obtained from these. Evidently, criterion validity is good-specific. The meta-analysis further highlighted the high levels of heterogeneity in the surveys, further questioning pooling of the results. An exploration of the heterogeneity through sub-group analysis did not yield any meaningful explanation as the reduction in some sub-groups can be explained by the lower numbers, not the rigour of the studies. Further investigation of the heterogeneity through meta-regression generated mixed results on potential drivers of the variation between stated and actual WTP values. This further questions the validity of estimates pooled across the different study settings and valuation goods. It is not yet clear from the literature how or whether the results can be transferred across settings and types of goods.

To help in interpreting and lending credibility to the responses and possibly also in forming adjustments that can enhance reliability, attempts should be made to collect additional data for cross tabulations (Arrow et al., 1993). Surveys should collect information on the respondent's background characteristics and socio-economic data such as income, attitudes towards the good or service and prior exposure or experience with the good. Such questions help in the interpretation of the primary valuation question and could also be used as further tests of validity of the data. The majority of reviewed comparisons did not report on the collection and use of such data in the assessment of criterion validity.

The review found a marked difference in the duration of time between surveys for hypothetical and actual values, with 65% occurring concurrently and 25% with more than a 4-week gap between the surveys. A two-week interval is the generally recommended retest period to enhance reliability of the values obtained (Duane, 1992). However, while longer durations could potentially introduce recall bias, short durations of time difference means that respondents may remember what they said in the hypothetical survey and deliberately repeat the value to appear publicly consistent. While a longer duration between the two surveys might offer the respondent sufficient time to think and possibly forget or change their original values, it also increases the possibility of real change occurring and thus justifying a change in any value given. The duration between the two surveys is likely to contribute to conclusions on the criterion validity of contingent values. In the meta-regressions, the ratio of stated and actual WTP values was higher when the two surveys were conducted at the same time. The duration between the two surveys was not identified as influencing hypothetical bias for studies presenting percent summaries.

The review also highlights some potential queries about how valid the comparisons of mean values were, not only raising questions of study quality but also how appropriate current conclusions might be. For example, 20% of comparisons did not include descriptions of how mean WTP/WTA was calculated, one-third of the comparisons had no information on tests used to determine differences in mean values between hypothetical and actual comparisons and there was a general absence of information on the treatment of missing values. There were also very few explanations given for the selection of value cues behind bid offers regardless of design format. Until there is a set of reliable reference surveys, the burden of proof of reliability and validity (of a CV Survey) rests on the survey designers and analysts (Neill et al., 1994; Onwujekwe et al., 2001). It is not clear what the impact of analytic methods has had on conclusions to date. In addition, poor reporting continues to limit the use of comparisons for systematic reviews and meta analyses in CV research (Trapero-Bertran et al., 2013). Queries on the methodological quality of comparisons also raises the broader issue of the potential for developing either an evidence-based set of guidelines for high quality WTP comparisons or appropriate reporting guidelines for CV comparisons.

Whilst the assessments are carried out in different sectors, the methods used to evaluate validity could be comparable and lessons transferred. With only a few comparisons identified, the health focussed comparisons seemed to use some appropriate methods compared with other sectors. For example, higher proportions of comparisons used the same respondents, administration modes, elicitation formats and payment vehicles in hypothetical and actual surveys. Comparisons also reported on key explanatory variables, allowing for a comparison within the sector and potential transferability of the methods used to assess criterion validity across sectors. Having the same respondent for the hypothetical and actual valuation scenarios reduces bias when judging criterion validity and this too occurred more frequently in the health focussed comparisons. However, the assessment of criterion validity could be enhanced in all sectors if values were elicited from comparisons with the closest relation to the planned intervention.

Appropriate estimation methods should be used and summary statistics provided in comparable formats, such as ratios. Ensuring content validity might also improve the tests. This can be achieved by conducting focus group discussions with key stakeholders in the valuation context. This would help achieve credible scenarios, determine suitable values for use in the surveys, appropriate study administration modes and payment vehicles. The payment vehicle forms a substantive part of the overall package under evaluation and is generally believed to be a non-neutral element of the survey (Bateman et al., 2002), affecting both the response rate and the magnitude of the values. The majority of the payment vehicles used in the surveys were amenable to a criterion validity assessment except coercive measures such as tax. It is difficult to assess how this payment vehicle was used in actual surveys and the results used to determine criterion validity.

## 5. Conclusions

The evidence on the criterion validity for CV comparisons is more mixed than authors are representing because substantial differences in study design between hypothetical and actual WTP/WTA surveys are not accounted for. This concern is compounded by the presence of key gaps in the reporting of methods and data. Infact, there does not seem to be a sufficient pool of criterion validity studies in sectors such as health, to permit a reasonable meta-synthesis and meta-analysis, and draw robust results.

Across the sectors, there was a general dearth of studies employing similar methods (e.g. WTP elicitation formats) combined with other attributes (such as respondent characteristics discussed earlier), to allow the testing of the effect of these on criterion validity. As a result, the evidence does not adequately support current conclusions on the criterion validity of WTP. Sufficient breadth of empirical evidence on the criterion validity of WTP across sectors and goods is needed. This would enable a dataset to facilitate more rigorous testing of the different experimental protocols that might influence the differences between stated and actual WTP values.

The WTP method offers potential for a welfare based measure of value for non-marketed goods and should not be subjected to the blanket criticism that it has received over the years based on findings from poorly designed comparisons that incorrectly suggest a lack of criterion validity. Such criticism might be one of the reasons for the slow uptake of the method in evaluations, especially in the health

sector. For evaluations of public health interventions where outcomes beyond quality-adjusted life years are often needed, this is particularly important. However, if the method is to be improved, more studies are required. The presented review and synthesis of the evidence further contributes to the ongoing work aimed at improving the method. The development of reporting guidelines for CV comparisons and the development of methodological guidelines for the conduct of criterion validity assessments would aid in the assessment of validity of the studies and transferability of findings.

## Conflict of interest

The authors declare no conflict of interest.

## Acknowledgments

This paper received financial support from Brunel University London, through PhD funding from the Health Economics Research Group.

## Appendix 1. Sample search strategy

EBSCOhost Interface. Databases: EconLit; CINAHL Plus

| # | Query | Results |
|---|---|---|
| S14 | S11 OR S12 OR S13 | 29 |
| S13 | S7 AND S10 | 0 |
| S12 | S6 AND S10 | 27 |
| S11 | S3 AND S10 | 2 |
| S10 | S8 OR S9 | 2393 |
| S9 | TI Will* AND Accept or WTA | 250 |
| S8 | TI Will* AND Pay or WTP | 2274 |
| S7 | S3 AND S6 | 2 |
| S6 | S4 AND S5 | 1229 |
| S5 | TI Actual OR revealed OR real OR inconsequentiality OR Direct | 46,914 |
| S4 | TI Stated OR Hypothetical OR Contingent OR Consequentiality OR indirect | 7109 |
| S3 | S1 AND S2 | 6622 |
| S2 | TI Validity OR Valid* | 33,026 |
| S1 | TI External OR Criterion OR Predictive OR Reliability | 34,557 |

## Appendix 2. Data extraction form items

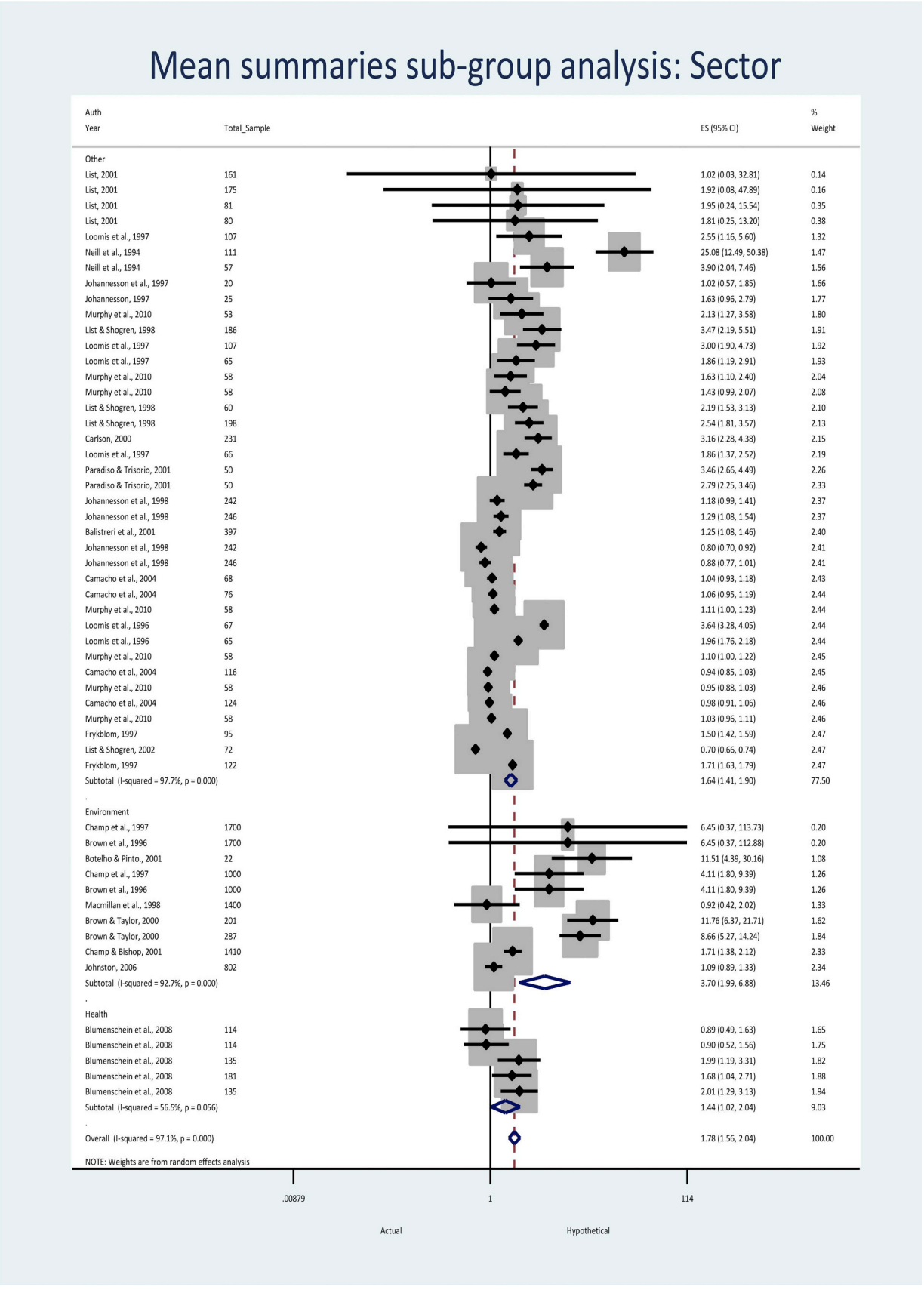
| General | Comments |
|---|---|
| Study Id | Sector |
| Study title | Good |
| Publication Year | Class of good |
| Study Country | Purpose of good |
| Study type | Validity term used |
| **Hypothetical and Actual surveys** | |
| Welfare measure | Payment duration |
| Study perspective | Study response rate (general) |
| Study technique | WTP response rate (WTP question) |
| Sample size | WTP estimation method |
| Sample type | Regression model used |
| Participation fee given | WTP summary given |
| Administration mode | WTP results (Mean/%ge) |
| Values elicitation format | WTP results (Median) |
| Bid values (where relevant) | WTP results (SD, SE, CI) |
| Payment vehicle | Statistical tests conducted and results |
| **Comparison between two studies** | |
| Respondent in both studies | Validity test results including ratios |
| Questionnaire used in both studies | Survey setting |
| Duration between surveys | Reasons given for disparity in hypothetical and actual values |
| Type of comparison (between/within) | Author conclusion on validity |
| Validity assessment method | |

## Appendix 3. Background characteristics of papers included in the review

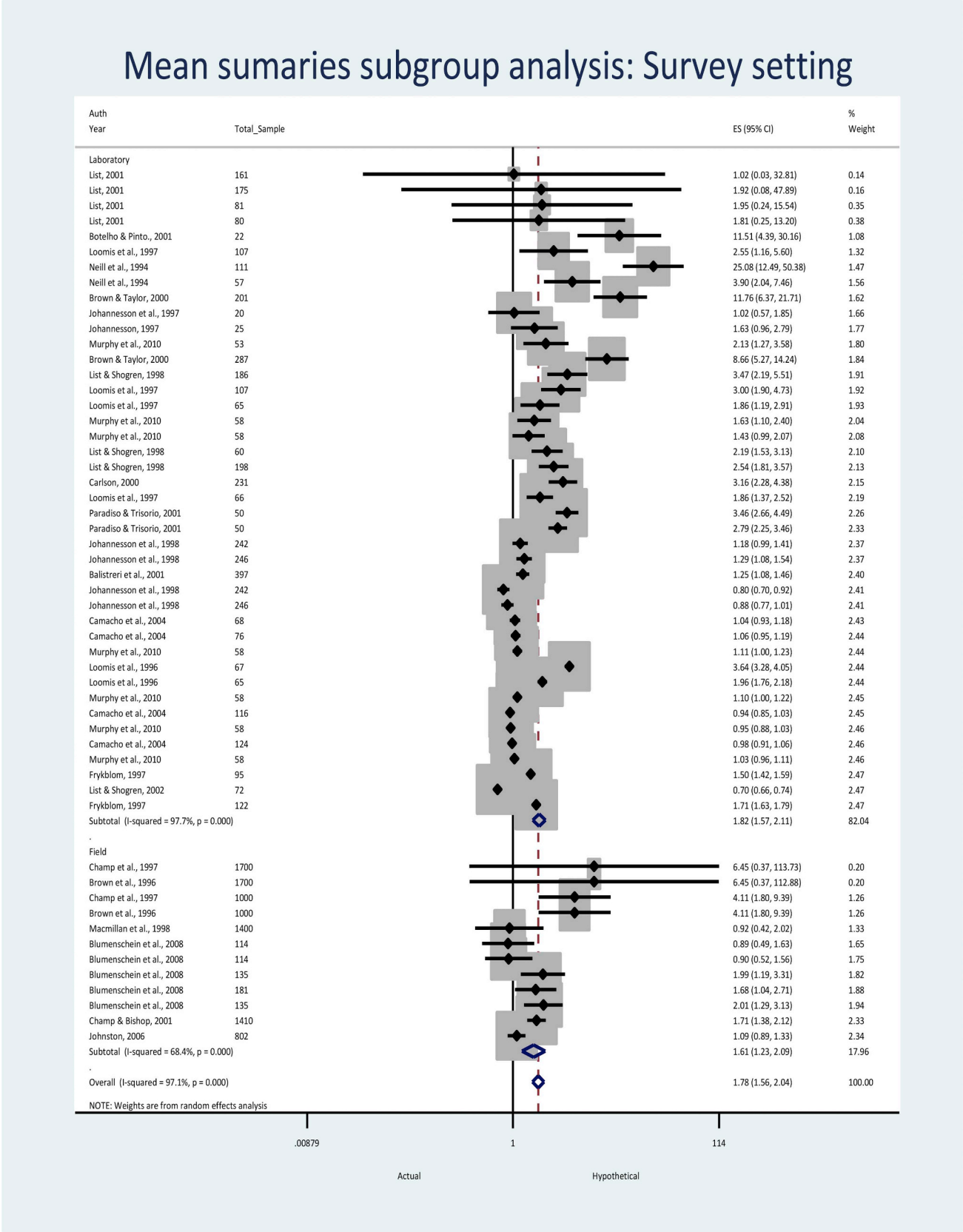
| No. | Reference | No. of comparisons | Country | Sector | Class of good | Validity term |
|---|---|---|---|---|---|---|
| 1 | Balistreri et al. (2001) | 1 | USA | Other | Pure Private | Non-Specific |
| 2 | Bhatia and Fox-Rushby (2003) | 1 | India | Health | Pure Private | Criterion |
| 3 | Bishop and Heberlein (1979) | 1 | USA | Environment | Quasi-Private | Non-Specific |
| 4 | (Blumenschein Blomquist, GC., Johannesson, M., Horn, N., Freeman, P 2008) | 6 | USA | Health | Pure Private | Non-Specific |
| 5 | (Blumenschein et al. 1998) | 2 | USA | Other | Pure Private | Hypothetical bias |
| 6 | Blumenschein et al. (2001) | 5 | USA | Health | Pure Private | Non-Specific |
| 7 | (Botelho and Pinto, 2002b) | 1 | Portugal | Environment | Quasi-Private | Hypothetical bias |
| 8 | Bratt (2010) | 3 | El Savador, Egypt | Health | Pure Private | Criterion |
| 9 | Brown et al. (1996) | 2 | USA | Environment | Pure Public | Non-Specific |
| 10 | Brown and Taylor (2000) | 2 | USA | Environment | Pure Public | Non-Specific |
| 11 | Bryan and Jowett (2010) | 1 | UK | Health | Pure Private | Hypothetical bias |

| | | | | | | |
|---|---|---|---|---|---|---|
| 12 | Byrnes et al. (1999) | 2 | USA | Environment | Quasi-Private | Non-Specific |
| 13 | Camacho-Cuena et al. (2004) | 4 | Spain | Mixed | Quasi-Private | Hypothetical bias |
| 14 | Carlson (2000) | 3 | USA | Other | Pure Private | Criterion |
| 15 | Champ and Bishop (2001) | 1 | USA | Environment | Quasi-Private | Non-Specific |
| 16 | Champ et al. (1997) | 2 | USA | Environment | Pure Public | Non-Specific |
| 17 | (Cummings Harrison, G.W., Rutstrom, E.R. 1995) | 5 | USA | Other | Pure Private | Non-Specific |
| 18 | Cummings et al. (1997) | 1 | USA | Environment | Quasi-Private | Non-Specific |
| 19 | Fox et al. (1998) | 2 | USA | Health | Pure Private | Non-Specific |
| 20 | Frykblom (1997) | 3 | Sweden | Other | Pure Private | Non-Specific |
| 21 | Getzner (2000) | 1 | Austria | Environment | Pure Public | Hypothetical bias |
| 22 | Heberlein and Bishop (1986) | 6 | USA | Other | Quasi-Private | Non-Specific |
| 23 | Johannesson (1997) | 1 | Sweden | Other | Pure Private | Hypothetical bias |
| 24 | Johannesson et al. (1997) | 1 | Sweden | Other | Pure Private | Non-Specific |
| 25 | Johannesson et al. (1998) | 4 | Sweden | Other | Pure Private | Non-Specific |
| 26 | Johnston (2006) | 1 | USA | Environment | Quasi-Private | Criterion |
| 27 | List (2001) | 4 | USA | Other | Pure Private | Hypothetical bias |
| 28 | List and Shogren (2002) | 1 | USA | Other | Pure Private | Non-Specific |
| 29 | List and Shogren (1998) | 3 | USA | Other | Pure Private | Non-Specific |
| 30 | (Loomis et al., 1996b) | 2 | USA | Other | Pure Private | Criterion |
| 31 | Loomis et al. (1997) | 4 | USA | Other | Pure Private | Non-Specific |
| 32 | Macmillan et al. (1999) | 1 | USA | Environment | Pure Public | Non-Specific |
| 33 | Mozumder and Berrens (2007) | 2 | USA | Other | Pure Public | Hypothetical bias |
| 34 | Muller and Ruffieux (2011) | 1 | France | Other | Pure Private | External |
| 35 | (Murphy Stevens, T., Weatherhead, D. 2002) | 4 | USA | Environment | Pure Public | Hypothetical bias |
| 36 | Murphy et al. (2010) | 9 | USA | Other | Pure Private | Hypothetical bias |
| 37 | Neill et al. (1994) | 2 | USA | Other | Pure Private | Non-Specific |
| 38 | Onwujekwe et al. (2001) | 6 | Nigeria | Health | Pure Private | Criterion |
| 39 | Onwujekwe and Uzochukwu (2004) | 2 | Nigeria | Health | Pure Private | Criterion |
| 40 | Onwujekwe (2004) | 3 | Nigeria | Health | Pure Private | Hypothetical bias |
| 41 | Onwujekwe et al. (2005) | 3 | Nigeria | Health | Pure Private | Criterion |
| 42 | Onwujekwe (2001) | 2 | Nigeria | Health | Pure Private | Predictive |
| 43 | Paradiso and Trisorio (2001) | 2 | UK | Other | Pure Private | Non-Specific |
| 44 | Ramke et al. (2009) | 3 | East Timor | Other | Pure Private | Criterion |
| 45 | (Seip Strand, J. 1992) | 1 | Norway | Environment | Pure Public | Non-Specific |
| 46 | Veisten and Navrud (2006) | 34 | Norway | Environment | Pure Public | Non-Specific |
| 47 | Vernazza et al. (2015) | 2 | UK, Germany | Health | Pure Private | Non-Specific |
| 48 | Vossler et al. (2003a,b) | 2 | USA | Environment | Quasi-Private | Criterion |
| 49 | Vossler and Kerkvliet (2003) | 3 | USA | Environment | Pure Public | Criterion |
| 50 | (K G Willis and Powe, 1998) | 1 | UK | Environment | Quasi-Private | Criterion |

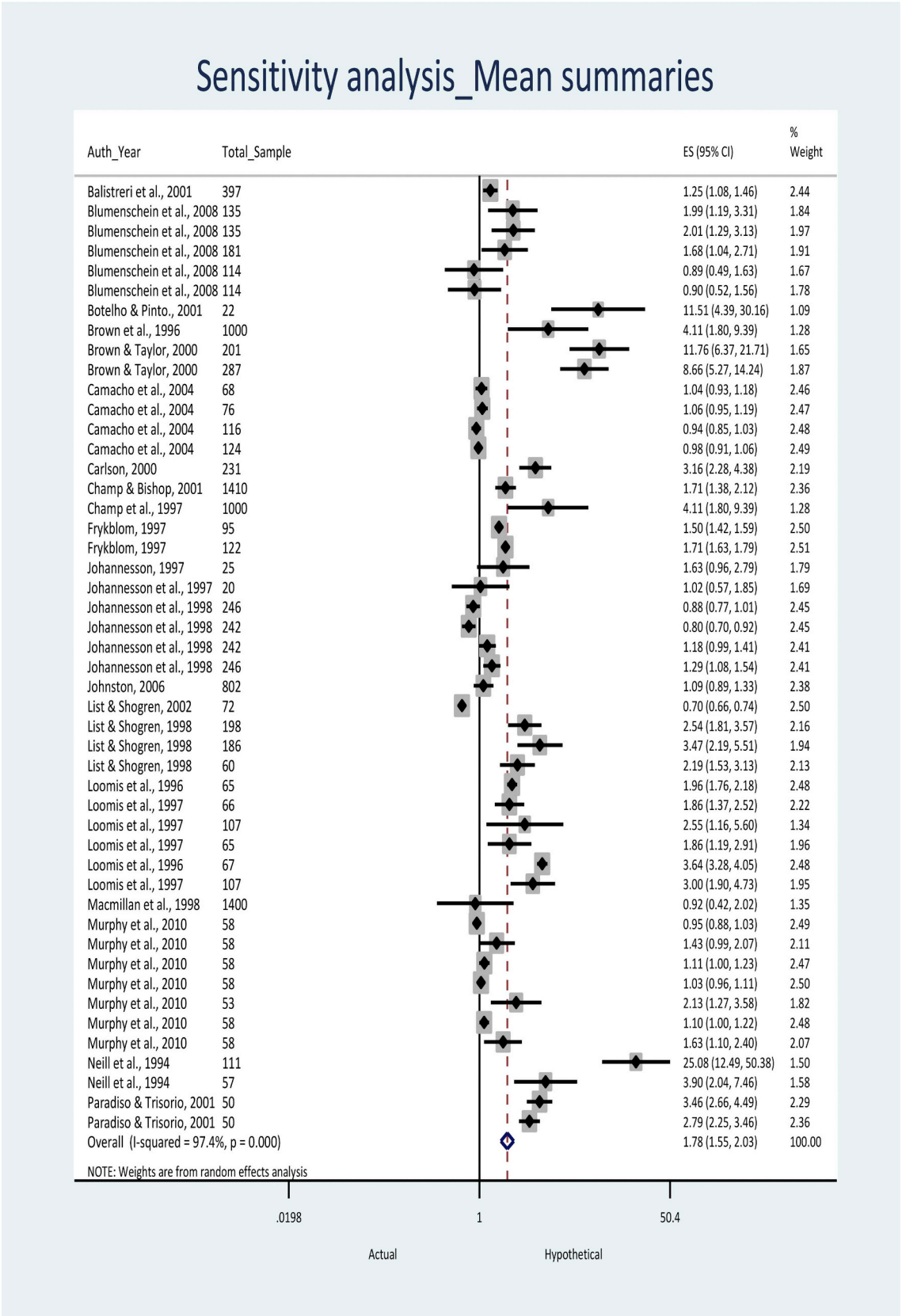
**Appendix 4. Subgroup analysis by sector for mean summaries**



**Appendix 5. Subgroup analysis by survey setting for mean summaries**

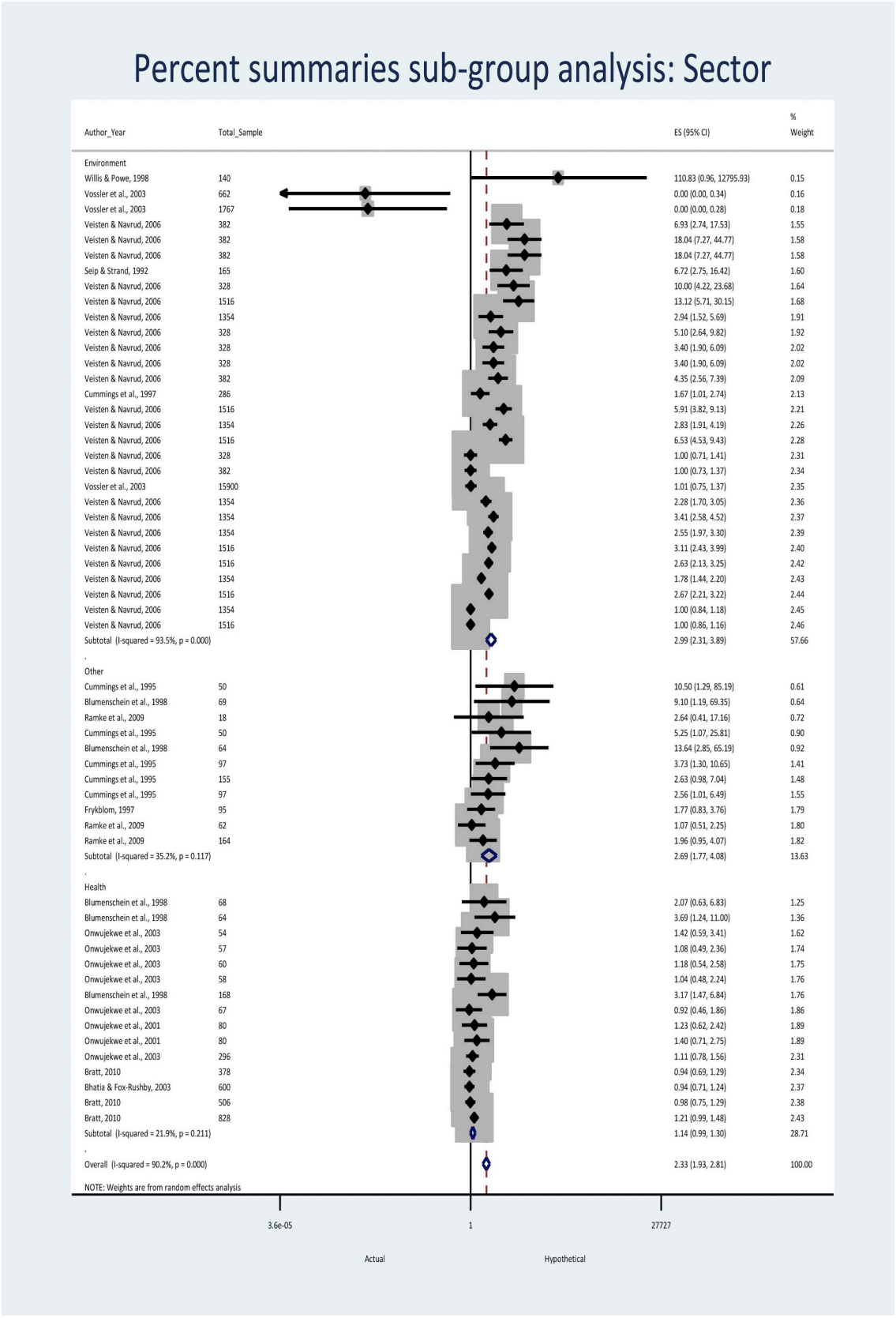


**Appendix 6. Sensitivity analysis for mean summaries**

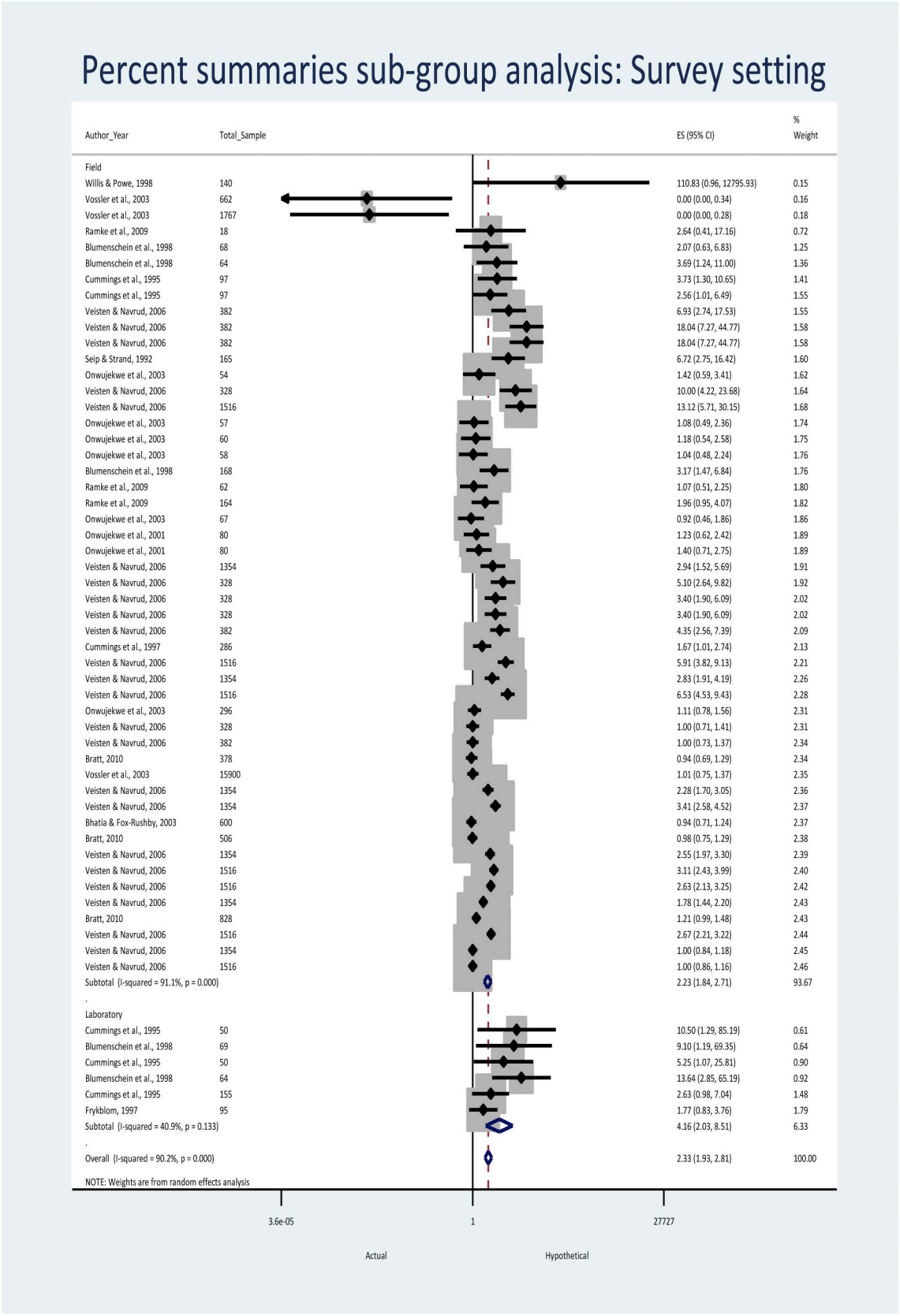



.

**Appendix 7. : Subgroup analysis by sector for percent summaries**

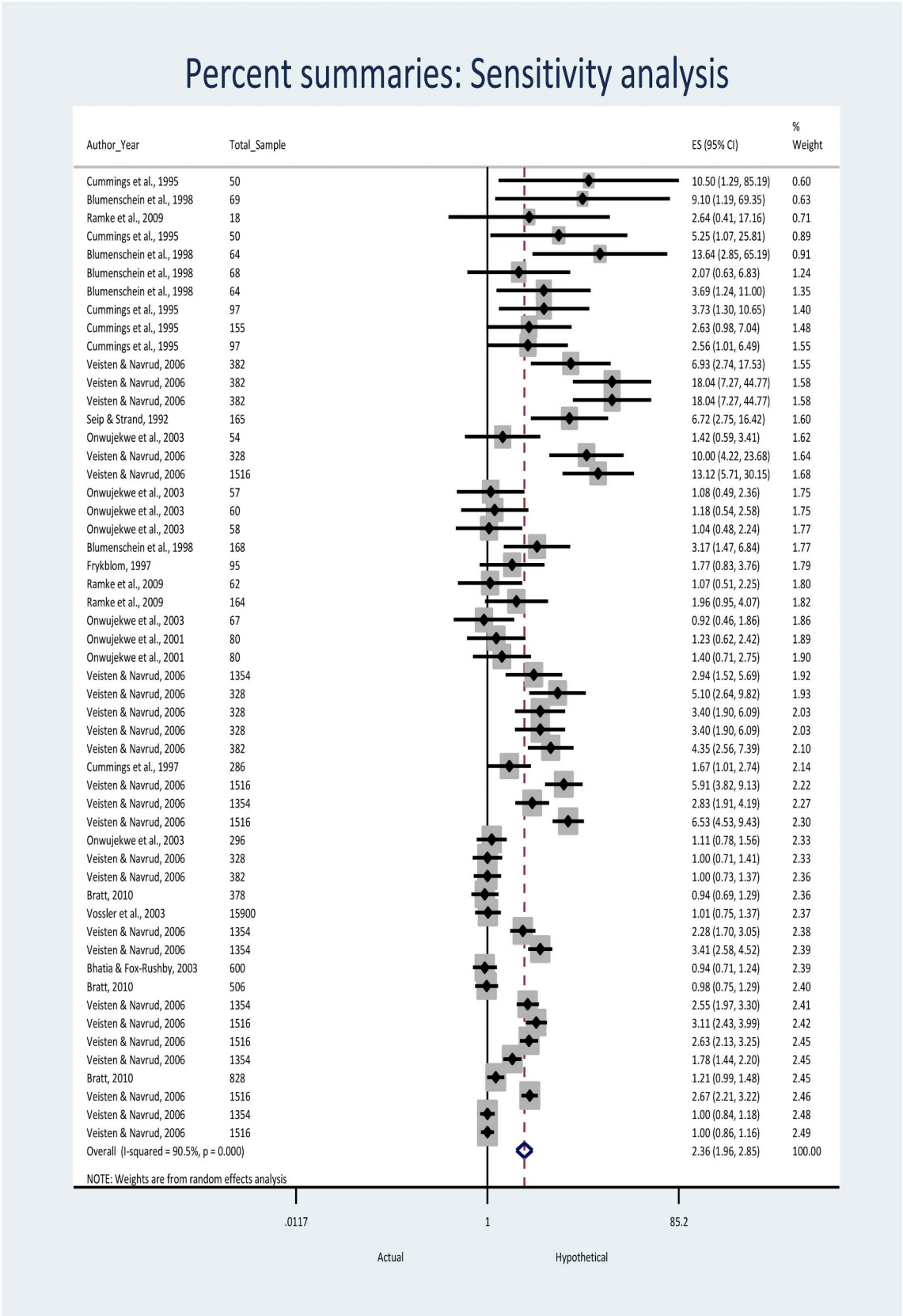


**Appendix 8. Subgroup analysis by survey setting for percent summaries**




.

**Appendix 9. Sensitivity analysis for percent summaries**




.

## References


Arrow, K., Solow, R., Portney, P.R., Leamer, E.E., Radner, R., Schuman, H., 1993. Report of the NOAA panel on contingent valuation. Fed. Regist. 58 (10), 4601–4614.

Balistreri, E., McClelelland, G., Poe, G., Schulze, W., 2001. Can hypothetical questions reveal true values? A laboratory comparison of dichotomous choice and open-ended contingent values with auction values. Environ. Resour. Econ. 18 (3), 275–292.

Bateman, I.J., Carson, R.T., Day, B., Hanemann, M., Hanleys, N., Hett, T., Jones-Lee, M., Loomes, G., Mourato, S., Ozdemiroglu, E., Pearce, D., Sugden, R., Swanson, J., 2002. Economic Valuation with Stated Preference Techniques: A Manual. Edward Elgar, Cheltenham (Cheltenham, UK).

Bhatia, M.R., Fox-Rushby, J.A., 2003. Validity of Willingness to Pay: hypothetical versus actual payment. Appl. Econ. Lett. 10 (12), 737–740.

Bishop, R.C., Heberlein, T.A., 1979. Measuring values of extramarket goods: are indirect measures biased? Am. J. Agric. Econ. 61 (5), 926–930.

Blumenschein, K., Johannesson, M., Blomquist, G., Liljas, B., O'Conor, R., 2014. Experimental results on expressed certainty and hypothetical bias in contingent valuation. South. Econ. J. 65 (1), 169–177.

Blumenschein, K., Johannesson, M., Yokoyama, K.K., Freeman, P.R., 2001. Hypothetical versus real willingness to pay in the health care sector: results from a field experiment. J. Health Econ. 20 (3), 441–457.

Blumenschein Blomquist, G.C., Johannesson, M., Horn, N., Freeman, P.,K., 2008. Eliciting willingness to pay without bias: evidence from a field experiment. Econ. J. 118, 114–137.

Botelho, A., Pinto, L.C., 2002. Hypothetical, real, and predicted real willingness to pay in open-ended surveys: experimental results. Appl. Econ. Lett. 9 (15), 993–996.

Bratt, J.H., 2010. Predicting impact of price increases on demand for reproductive health services: can it be done well? Health Policy 95 (2–3), 159–165.

Brown, K.M., Taylor, L.O., 2000. Do as you say, say as you do: evidence on gender differences in actual and stated contributions to public goods. J. Econ. Behav. Organ. 43 (1), 127–139.

Brown, T.C., Champ, P.A., Bishop, R.C., McCollum, D.W., 1996. Which response format reveals the truth about donations to a public good? Land Econ. 72 (2), 152–166.

Bryan, S., Jowett, S., 2010. Hypothetical versus real preferences: results from an opportunistic field experiment. Health Econ. 19, 1502–1509.

Byrnes, B., Jones, C., Goodman, S., 1999. Contingent valuation and real economic commitments: evidence from electric utility green pricing programmes. J. Environ. Plan. Manag. 42 (2), 149–166.

Camacho-Cuena, E., García-Gallego, A., Georgantzis, N., Sabater-Grande, G., 2004. An experimental validation of hypothetical WTP for a recyclable product. Environ. Resour. Econ. 27 (3), 313–335.

Cameron, A.C., Trivedi, P.K., 2009. Microeconometrics Using Stata. Stata Pres, Texas.

Carlson, J., 2000. Hypothetical surveys versus real commitments: further evidence. Appl. Econ. Lett. 7 (7), 447–450.

Carmines, E.G., Zeller, R.A., 1979. Reliability and Validity Assessment. Sage Publications, Beverly Hills; London.

Carson, R.T., FLores, N.E., Martin, K.M., Wright, J.L., 1996. Contingent valuation and revealed preference methodologies: comparing the estimates for quasi-public goods. Land Econ. 72 (1), 80–99.

Champ, P.A., Bishop, R.C., Brown, T.C., McCollum, D.W., 1997. Using donation mechanisms to value nonuse benefits from public goods. J. Environ. Econ. Manag. 33 (2), 151–162.

Champ, P.A., Bishop, R.C., 2001. Donation payment mechanisms and contingent valuation: an empirical study of hypothetical bias. Environ. Resour. Econ. 19 (4), 383–402.

Cummings, R.G., Elliot, S., Harrison, G.W., Murphy, J., 1997. Are hypothetical referenda incentive compatible? J. Political Econ. 105 (3), 609–621.

Cummings, R.G., Harrison, G.W., Rutstrom, E.E., 1995. Homegrown values and hypothetical surveys: is the dichotomous choice approach incentive-compatible? Am. Econom. Assoc. 85 (1), 260–266.

Donaldson, C., Shackley, P., 2002. Willingness to pay for health care. In: Scott, A. (Ed.), Advances in Health Economics. John Wiley and Sons, Chichester, UK.

Duane, F.A., 1992. Information transmission in the survey interview: number of response categories and the reliability of attitude measurement. In: In: Alwin, Duane F. (Ed.), Source : Sociological Methodology Vol.22 ( 1992 ), pp . 83-118 Published by : American Sociologi. Socilogy and methodology, 22(May), pp.83–118.

Fox, J.A., Shogren, J.F., Hayes, D.J., Kliebenstein, J.B., 1998. CVM-X: calibrating contingent values with experimental auction markets. Am. J. Agric. Econ. 80 (3), 455–465.

Frykblom, P., 1997. Hypothetical question modes and real willingness to pay. J. Environ. Econ. Manag. 34 (3), 275–287.

Getzner, M., 2000. Hypothetical and real economic commitments, and social status, in valuing a species protection programme. J. Environ. Plan. Manag. 43 (4), 541–559.

Hanley, N., Splash, C., 1993. Cost-benefit Analysis and the Environment. E. Elgar.

Harrison, G.W., Rutström, E.E., 2008. Experiemental evidence on the existence of hypothetical bias in value elicitation methods. Handbook of Experimental Economics Results. pp. 752–767.

Heberlein, T.A., Bishop, R.C., 1986. Assessing the Validity of Contingent Valuation: Three Field Experiements, vol.56. pp. 99–107.

Higgins, J.P.T., Thompson, S.G., Deeks, J.J., Altman, D.G., 2003. Measuring inconsistency in meta-analyses. Br. Med. J. 327 (7414), 557–560.

Hosmer, D., Lemeshow, S., 2013. Applied Logistic Regression. Wiley Blackwell, New York.

Johannesson, M., 1997. Some further experimental results on hypothetical versus real willingness to pay. Appl. Econ. Lett. 4 (8), 535–536.

Johannesson, M., Liljas, B., Johansson, P.-O., 1998. An experimental comparison of dichotomous choice contingent valuation questions and real purchase decisions. Appl. Econ. 30 (5), 643–647.

Johannesson, M., Liljas, B., O'Conor, R.M., 1997. Hypothetical versus real willingness to pay: some experimental results. Appl. Econ. Lett. 4 (3), 149–151.

Johnston, R.J., 2006. Is hypothetical bias universal? Validating contingent valuation responses using a binding public referendum. J. Environ. Econ. Manag. 52 (1), 469–481.

Klose, T., 2003. A utility-theoretic model for QALYs and willingness to pay. Health Econ. 12 (1), 17–31.

Kealy, M.J., Montgomery, M., Dovidio, J.F., 1990. Reliability and predictive validity of contingent values: does the nature of the good matter? J. Environ. Econ. Manag. 19 (3), 244–263.

Liljas, B., Blumenschein, K., 2000. On hypothetical bias and calibration in cost–bene t studies. Health Policy 52, 53–70.

List, J.A., 2001. Do explicit warnings eliminate the hypothetical bias in elicitation procedures? Evidence from field auctions for sportscards. Am. Econ. Rev. 91 (5), 1498–1507.

List, J.A., Gallet, C.A., 2001. What experimental protocol influence disparities between actual and hypothetical stated values? Environ. Resour. Econ. 20 (3), 241–254.

List, J.A., Shogren, J.F., 1998. Calibration of the difference between actual and hypothetical valuations in a field experiment. J. Econ. Behav. Organ. 37 (2), 193–205.

List, J.A., Shogren, J.F., 2002. Calibration of willingness-to-accept. J. Environ. Econ. Manag. 43 (2), 219–233.

Little, J., Berrens, R., 2003. Explaining disparities between actual and hypothetical stated values: further investigation using meta-analysis. Econ. Bull. 3 (1).

Lipsey, M.W., Wilson, D.B., 2001. Practical Meta-Analysis. Sage, London.

Loomis, J., Bell, P., Cooney, H., Asmus, C., 2009. A comparison of actual and hypothetical willingness to pay of parents and non-parents for protecting infant health: the case of nitrates in drinking water. J. Agric. Appl. Econ. 41 (3), 697.

Loomis, J., Brown, Y., Lucero, B., Peterson, G., 1997. Evaluating the validity of the dichotomous choice question format in contingent valuation. Environ. Resour. Econ. 10 (2), 109–123.

Loomis, J., Brown, T., Lucero, B., Peterson, G., 1996. Improving validity experiments of contingent valuation methods: results of efforts to reduce the disparity of hypothetical and actual willingness to pay. Land Econ. 72 (4), 450–461.

Macmillan, D.C., Smart, T.S., Thorburn, A.P., 1999. A field experiment involving cash and hypothetical charitable donations. Environ. Resour. Econ. 14 (3), 399–412.

Mitchell, R.C., Carson, R.T., 1989. Using Surveys to Value Public Goods: the Contingent Valuation Method. Resources for the future, Washington DC.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D.G., The PRISMA Group, 2009. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS Med. 6 (7).

Mozumder, P., Berrens, R.P., 2007. Investigating hypothetical bias: induced-value tests of the referendum voting mechanism with uncertainty. Appl. Econ. Lett. 14 (10), 705–709.

Muller, L., Ruffieux, B., 2011. Do price-tags influence consumers' willingness to pay? On the external validity of using auctions for measuring value. Exp. Econ. 14 (2), 181–202.

Munro, A., 2009. Introduction. In: Bounded Rationality and Public Policy. Springer Netherlands, Dordrecht, pp. 1–16.

Murphy, J., Allen, G.P., Stevens, T.H., Weatherhead, D., 2004. A Meta-Analysis of Hypotheticalbias in Stated Preference Valuation. Department of Resource Economics. *University Of massachusetts*.

Murphy, J.J., Stevens, T.H., Yadav, L., 2010. A comparison of induced value and homegrown value experiments to test for hypothetical bias in contingent valuation. Environ. Resour. Econ. 47 (1), 111–123.

Murphy Stevens, T., Weatherhead, D.J.J., 2002. An Empirical Study of Hypothetical Bias in Voluntary Contribution Contingent Valuation: DOes Cheap Talk Matter?, Congress P.

Neill, H.R., Cummings, R.G., Ganderton, P.T., Harrison, G.W., McGukin, T., 1994. Hypothetical surveys and real economic commitments. Land Econ. 70 (2), 145–154.

Onwujekwe, O., 2004. Criterion and content validity of a novel structured haggling contingent valuation question format versus the bidding game and binary with follow-up format. Soc. Sci. Med. 58 (3), 525–537.

Onwujekwe, O., Chima, R., Shu, E., Nwagbo, D., Okonkwo, P., 2001. Hypothetical and actual willingness to pay for insecticide-treated nets in five Nigerian communities. Trop. Med. Int. Health 6 (7), 545–553.

Onwujekwe, O., 2001. Searching for a better willingness to pay elicitation method in rural Nigeria: the binary question with follow-up method versus the bidding game technique. Health Econ. 10 (2), 147–158.

Onwujekwe, O., Hanson, K., Fox-Rushby, J., 2005. Do divergences between stated and actual willingness to pay signify the existence of bias in contingent valuation surveys? Soc. Sci. Med. 60 (3), 525–536.

Onwujekwe, O., Uzochukwu, B., 2004. Stated and actual altruistic willingness to pay for insecticide-treated nets in Nigeria: validity of open-ended and binary with follow-up questions. Health Econ. 13 (5), 477–492.

Paradiso, M., Trisorio, A., 2001. The effect of knowledge on the disparity between hypothetical and real willingness to pay. Appl. Econ. 33 (11), 1359–1364.

Ramke, J., Palagyi, A., du Toit, R., Brian, G., 2009. Stated and actual willingness to pay for spectacles in timor-leste. Ophthalmic Epidemiol. 16 (4), 224–230.

Seip, K., Strand, J., 1992. Willingness to Pay for Environmental Goods in Norway: A Contingent Valuation Study with Real Payment. Working pa.

Slothuus, U., 2000. Economic Evaluation : theory , methods & application. Health Econom. Papers 5.

Slothuus, U., Larsen, M.L., Junker, P., 2002. The contingent ranking method - a feasible and valid method when eliciting preferences for health care? Soc. Sci. Med. 54 (10),

1601–1609.
Smith, R., Olsen, J.A., Harris, A., 1999. A Review of Methodological Issues in the Conduct of Willingness-To-Pay Studies in Health Care I: Construction and Specification of the Contingent Market.
Spencer, M.A., Swallow, S.K., Miller, C.J., 1998. Valuing water quality monitoring: a contingent valuation experiment involving hypothetical and real payments. Agric. Res. Econom. Rev. 27, 28–42.
StataCorp, 2015. Stata Statistical Software: Release, vol. 14 StataCorp LP, College Station, TX.
Trapero-Bertran, M., Mistry, H., Shen, J., Fox-Rushby, J., 2013. A systematic review and meta-analysis of willingness-to-pay values: the case of malaria control interventions. Health Econ. 22 (4), 428–450.
Veisten, K., Navrud, S., 2006. Contingent valuation and actual payment for voluntarily provided passive-use values: assessing the effect of an induced truth-telling mechanism and elicitation formats. Appl. Econ. 38 (7), 735–756.
Vernazza, C.R., Wildman, J.R., Steele, J.G., Whitworth, J.M., Walls, A.W.G., Perry, R., Mathews, P.H., Donaldson, C., 2015. Factors affecting patient valuations of caries prevention: using and validating the willingness to pay method. J. Dent. 43 (8), 981–988.
Veronesi, M., Alberini, A., Cooper, J.C., 2011. Implications of bid design and willingness-to-pay distribution for starting point bias in double-bounded dichotomous choice contingent valuation surveys. Environ. Resour. Econ. 49 (2), 199–215.
Vossler, C.A., Kerkvliet, J., Polasky, S., Gainutdinova, O., 2003a. Externally validating contingent valuation: an open-space survey and referendum in Corvallis, Oregon. J. Econ. Behav. Organ. 51 (2), 261–277.
Vossler, C.A., Ethier, R.G., Poe, G.L., Welsh, M.P., 2003b. Payment certainty in discrete choice contingent valuation responses: results from a field validity test. South. Econ. J. 69 (4), 886–902.
Vossler, C.A., 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 (3), 631–649.
Vossler, C.A., Watson, S.B., 2013. Understanding the consequences of consequentiality: testing the validity of stated preferences in the field. J. Econ. Behav. Organ. 86, 137.
Willis, K.G., Powe, N.A., 1998. Contingent valuation and real economic commitments: a private good experiment. J. Environ. Plan. Manag. 41 (5), 611–619.