Investigating the criterion validity of contingent valuation-willingness to pay methods

A Thesis Submitted for the Degree of Doctor of Philosophy

By

Gladys Lucy Wanjiru Kanya

Clinical Sciences Department, Brunel University London

2018

Abstract

With theoretical foundations in welfare theory, the cost benefit analysis (CBA) technique is a powerful tool for assessing benefits particularly where markets do not exist or would fail (for example due to the existence of public goods) or have become potentially politically excluded (such as the health sector). Unlike other economic evaluation techniques, costs and benefits are measured in monetary terms allowing for comparisons within and between different sectors of the economy for resource allocation decisions. Using contingent valuation (CV) techniques, people's preferences for goods are determined by finding out what they would be willing to pay (WTP) for specified benefits or improvements; or accept (WTA), as compensation for withdrawal or loss of benefit. While the use of WTP methods has grown in other sectors, the uptake in health has been limited. A long-standing criticism is that stated WTP estimates may be poor indicators of actual WTP, calling into question their validity and the use of such estimates for welfare valuation. The aim of this thesis is to investigate the criterion validity of CV-WTP studies. A four-pronged approach including critical appraisals of the available literature and evidence on criterion validity and empirical analyses was adopted. The thesis established the scarcity in criterion validity assessments, particularly in the health sector and that evidence on the criterion validity of CV-WTP is more varied than authors are presenting. The variety in the methods used to assess and report criterion validity assessments is demonstrated. Further, the impact of the analysis of hypothetical WTP on criterion validity assessments and conclusions thereof is demonstrated. The empirical analyses further demonstrate the differences in predictions and predictors of WTP analyses, discussing the effect of these on criterion validity assessments and conclusions. Finally, the thesis offers suggestions for the reporting of criterion validity assessments, in efforts to improve the method.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

CBA	Cost Benefit Analysis
CEA	Cost Effectiveness Analysis
CUA	Cost Utility Analysis
CV	Contingent Valuation[1]
CV	Compensating Variation[2]
CS	Compensating Surplus
DALY	Disability Adjusted Life Years
EV	Equivalent Variation
ES	Equivalent Surplus
HYE	Healthy Years Equivalent
NICE	National Institute for Health and Care Excellence
NOAA	National Oceanic and Atmospheric Administration
QALY	Quality Adjusted Life Years
TEV	Total Economic Value
TTO	Time Trade Off
WTP	Willingness to Pay
WTA	Willingness to Accept
UK	United Kingdom
USA	United States of America

[1] Benefit valuation technique

[2] Welfare measure

Acknowledgements

All the glory and honor to the Almighty God for His love and sufficient grace always, and for the lessons He has taught me through this journey: Faith, Trust, Obedience.

I would like to express my most sincere gratitude to my supervisors, Dr. Nana Anokye and Dr. Subhash Pokherel; for their unwavering support, patience, motivation and enthusiasm in guiding me throughout the development of this thesis.

I wish to acknowledge the receipt of a scholarship from Brunel University, London, through the Health Economics Research Group (HERG). I am very grateful too to earlier members of my supervisory team for their support: Professor Julia Fox-Rushby, Dr. Sabina Sanghera, Dr. Alex Lewin and Professor Joanne Lord.

My children, Miriam and Moses, my forever cheerleaders; you are the wind beneath my wings; every time I count my blessings I count you many times over! Thank you for enduring my lengthy periods of absence, physical or otherwise, holding me up in prayer and every support you could avail. My parents, Margaret & Peter Kanya; you always believed in me; your love, prayers, moral & emotional support upheld me. My siblings Robert, Grace, & Benjamin; thanks a bunch for all! My Cucus Rosemary Wambui, Rosaita Nzyoka & Grace Muthike, special family friends, Agnes & John Muli, Majors David & Magdalene Mutokaa, Agnes & Joseph Kyungu; thank you for sending me off on this journey with love & prayers; heaven heard your intercessory prayers for me all the time.

To my very special ladies: Wendy Malombe, Janet Munyasya, Irene Muembu, Kanyiva Muindi, Florence Thungu-Maloba and Agnes Kisese; I count myself lucky to have you in my life, thank you for your unwavering support always. To my friends at SA Hillingdon Corps; and, John and Hazel Savil, your thoughts, prayers and encouragement through this journey are greatly appreciated.

Finally, my most profound gratitude to a special sister and friend, Dr. Carol Dayo Obure; your unwavering love and support kept me going, and out of the m-house. Thank you for believing in me, managing me and mine, enduring many lengthy calls, chats, ch-moments and all with such grace. You are an angel and I thank God for blessing me with you. .

Dedication

This work is dedicated with fervent love

to my children Miriam and Moses;

and my parents Margaret and Peter Kanya.

"I have held many things in my hands, and I have lost them all; but whatever I have placed in God's hands, that, I still possess." –Corrie ten Boom

Chapter 1 Introduction

1.1 Background

Estimating the costs and associated benefits of interventions forms a critical information base for supporting the design and decisions inherent in programming and resource planning. Economic evaluations are conducted to assign values to the costs, outcomes, consequences or choices made on policies, regulations, projects and programmes. Such evaluations include comparisons of different measures of efficiency, costs and benefits to determine the most suitable for particular situations. Depending on factors such as the economic question, outcome measure of interest, and whether costs and benefits are valued in monetary terms, techniques such as cost-effectiveness[3] (CEA), cost-utility[4] (CUA), and cost benefit analyses (CBA) are employed for this purpose (Drummond et al. 1987).

Unlike the CEA and CUA techniques which measure program or intervention outcomes in non-monetary units such as quality adjusted life years (QALYs); the cost benefit analysis (CBA) technique measures both costs and outcomes in monetary units, allowing for comparisons of programmes both within and between different sectors of the economy and thus aiding resource allocation decisions[1]. Further, while CEA and CUA analyses are based on a "decision making" philosophy where elected or appointed decision-makers review results and decide on the relative values assigned to competing programmes and goals; the philosophical foundation of the CBA technique is in the principles of welfare economics where the relevant source of values is believed to be individual consumers (Sugden & Williams 1978). A key premise of this technique therefore is that individual consumers are deemed to be the ideal source of monetary values for programme costs and outcomes.

[3] Cost-effectiveness analysis (CEA): A full economic evaluation where both the costs and consequences of interventions, programs or policies are examined. In this, costs are related to a single, common effect (such as life years gained) that may differ in magnitude between alternate programmes (Drummond, 2005). The results are expressed as cost per measure of effect.

[4] Cost-utility analysis (CUA): A full economic evaluation where the incremental cost of a programme from a particular viewpoint is compared to the incremental health improvement attributable to the programme; and where the health improvement is measured in quality-adjusted life years (QALYs) gained, or some variant like disability-adjusted life-years (DALYs) gained. The results are expressed as a cost per QALY gained (Drummond, 2005).

In aggregating the values of individuals, attempts are made to measure the ‘welfare’ of society as a whole. In this measurement, benefits are defined as increases in utility and costs are defined as reductions in utility. In fields such as environment, the assessment of the benefits or damages gained or lost from choices such as the implementation of a policy have been demonstrated using CBA analyses (Michell & Carson 1989). Further, the method allows for the scale of these damages or benefits to be demonstrated. In one of the earliest documented arbitrated case of damages to natural resources, the US department’s National Oceanic and Atmospheric Administration (NOAA) commissioned a panel of experts to assess economic value of the damage caused from an oil spillage (Arrow et al. 1993). The team recommended the use of CBA techniques in this economic evaluation. In the UK, regulatory impact assessments of national and European community directives often employ monetary valuation techniques in the assessment of both marketed and non-marketed goods (Bateman et al. 2002).

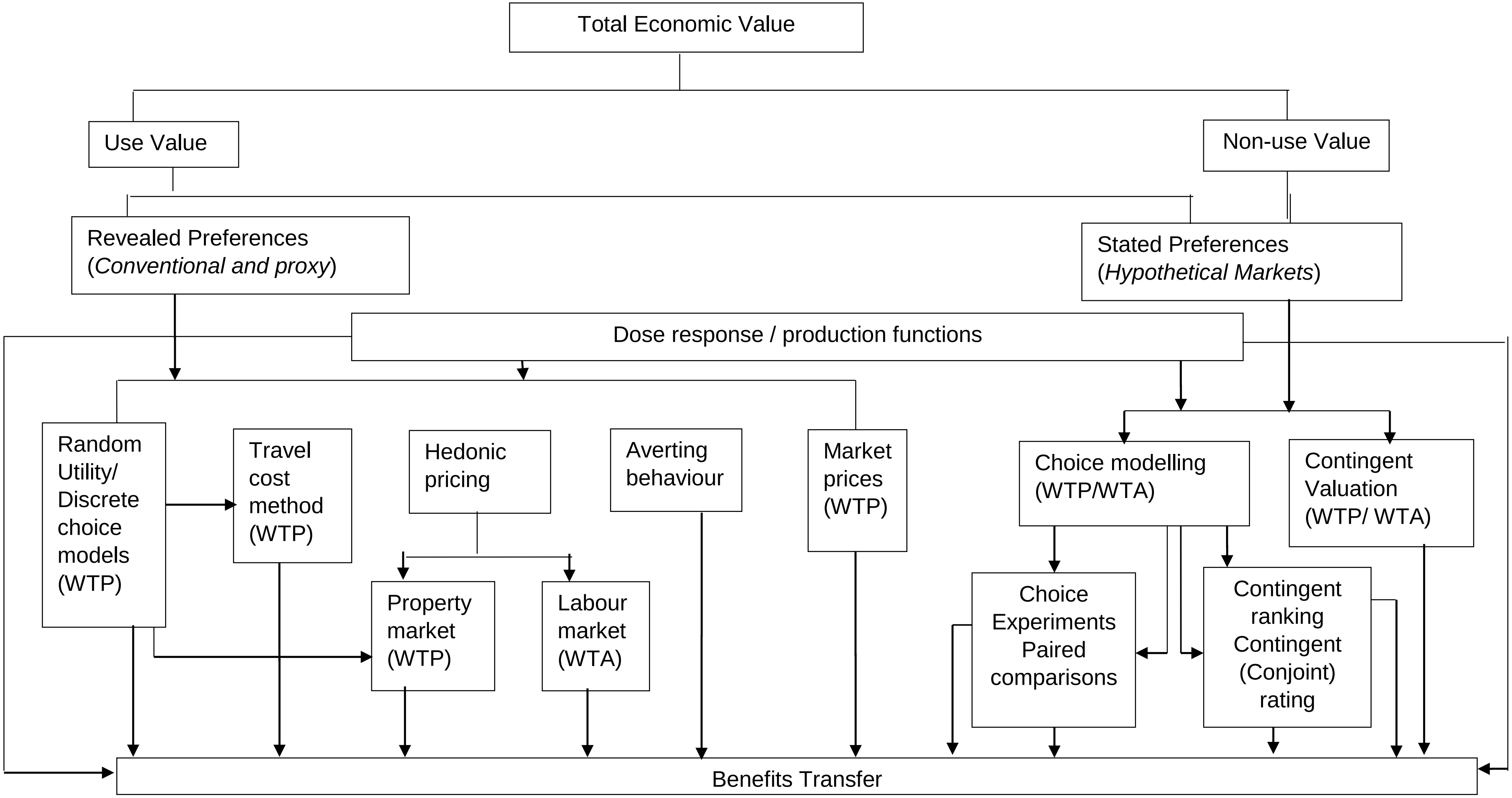
A crucial feature of economic valuation is that there are no absolute measures of value, only equivalences between one thing (commodity or service) and the other (Bateman et al. 2002). Benefits and costs are defined in terms of each other in this way: the measure of the benefit that an individual receives from something is how much he or she is willing to give up to obtain it and the more the individual is willing to give up, the larger the supposed benefit derived from the good. On the other hand, to measure how much it costs one to give up something then we measure how much the individual would be willing to be compensated in exchange for the good and the higher the compensation the individual would want the higher the cost of the good. Economic valuation does not take any position about what would be considered good for people, using instead the relative valuations as are revealed in people’s preferences to judge the benefit and costs placed on the goods (Bateman & Turner 1993). For example, in the valuation of natural resources such as is presented the Arrow (1993) case, the equivalences would be the ocean as it was before and following the oil spillage. The economic concept of preferences is therefore used to define benefits and costs (Gafni, 1991).

Economic evaluation involves the determination of both use and non-use values. The term “use values” refers to the value placed on or experienced by individuals or communities who, in a variety of ways, make active use of the resources being valued (Arrow et al. 1993). The measurement of use values is relatively easy and is done through the information revealed in market transactions, such as the demand and prices paid for the goods or services (Bateman et al. 2002). A second set of values commonly assessed in economic evaluations are non-use values. These refers to the value placed on a resource by individuals who do not actively or immediately use the resources being valued but derive satisfaction from the mere existence of the resource, even if they never intend to make use of the resource in the foreseeable future (Bateman et al. 2002). Such would suffer or feel “loss” if such a resource were to disappear (Bateman et al. 2002). These values are also referred to as “existence” (letting these resources exist in their own right, for instance a forest reserve) or “passive” values. Conservation of such resources may also be supported for the purpose of passing on to future generations[5] (Bateman et al. 2002). Classified in these values also is altruistic values, where individuals derive satisfaction from availing a resource or good to another person where the value of the good or resource may not directly benefit the person paying for it (Bateman et al. 2002). The valuation of non-use values is more complicated as the resource involved is not directly traded in the market.

According to the theory of welfare economics, an economy is in an equilibrium state when resources are allocated in such a way that no further gains of economic efficiency are possible (Bateman & Turner 1993; Carson 2012; Bateman et al. 2002; Mitchell & Carson 1989; Carson 2000). The accurate determination of the economic efficiency of a program or policy decision necessitates the determination of the total economic value (TEV). The TEV includes both the use and non-use values. Figure 1.1 illustrates the techniques used to measure the concepts of use and non-use values.

[5] Bequest value such as a species of wild animals or a park.

Figure 1-1: Economic evaluation techniques


Total Economic Value
Use Value
Non-use Value
Revealed Preferences (*Conventional and proxy*)
Stated Preferences (*Hypothetical Markets*)
Dose response / production functions
Random Utility/ Discrete choice models (WTP)
Travel cost method (WTP)
Hedonic pricing
Averting behaviour
Market prices (WTP)
Choice modelling (WTP/WTA)
Contingent Valuation (WTP/ WTA)
Property market (WTP)
Labour market (WTA)
Choice Experiments Paired comparisons
Contingent ranking Contingent (Conjoint) rating
Benefits Transfer

Source: (Bateman et al. 2002)

These include market prices, averting behaviour techniques and travel cost methods, all of which employ the WTP method. Hedonic pricing techniques which employ property market (WTP) and labour market techniques (WTA) are also used to determine use values. Finally, random utility or discrete choice models which employ WTP are also used. These techniques are explained further in Appendix 1. Use values can also be determined using stated preference techniques such as choice modelling and contingent valuation (CV). This thesis specifically focusses on the use of CV methods and these are discussed further in the next section.

The contingent valuation (CV) method draws upon economic theory and the methods of survey research to elicit directly from people the values they place upon goods for which a market does not exist (Bateman et al. 2002). These are also referred to as their stated preferences (SP). The method has primarily been used to elicit preferences for public and quasi-public goods (Mitchel & Carson 1989). Survey questions are used to elicit people's preferences for goods by finding out what they would be willing to pay (WTP) for specified benefits or improvements in them. Respondents may also be asked to state the level of compensation they would be willing to accept (WTA) for a given withdrawal or loss of a benefit. In this way, the method elicits individual's WTP in dollar amounts (Mitchell & Carson 1989). In either valuation, consumers are presented with a hypothetical market in which they have an opportunity to buy the good in question and the elicited WTP values are therefore contingent upon the hypothetical market described to the respondent. The hypothetical market may be modelled after either a private or public good.

Willingness to pay, the technique used in CV studies, has a formal relationship to the notion of a demand curve (Bateman et al. 2002). Like the theory of demand, in engaging in a WTP transaction, the theory of consumer behaviour is assumed (Drèze & Stern 1987). In this, the consumer is assumed to derive satisfaction (utility) from the consumption of the valuation good; that this utility is connected to his preferences and that the consumer can order or rank their preferences in a rational manner to obtain the maximum satisfaction (Varian 2006). In ranking their preferences, it is assumed that the consumer applies the properties of preferences namely completeness, transitivity, non-satiation and reflexivity (Hardwick et al. 1986). Demand theory further suggests that a consumer will consume more of a good at a lower price and vice versa (Hardwick et al. 1986; Varian 2014).

However, the market price for a commodity does not always match the maximum amount of money that the consumer is willing to pay for this good. The difference between the maximum amount of money that a consumer is willing to pay for a good and the market price for the good is known as consumer surplus (Varian 2014). By employing appropriate elicitation techniques, we can determine the consumer surplus using the CV method. The CV method has been used extensively in the environment sector but less so in other sectors such as the health sector. Critiques of the method cite, among other issues, the hypothetical bias inherent in CV studies. As a result, the CEA technique is preferred, with the quality of life adjusted years (QALY), used as the measure of effect. In the next section, the use of the CEA technique in health is further explored. A critique of this method is provided and the CV-WTP, offered as an alternative, is presented. The section concludes with a discussion on the challenges to the widespread adoption of the CV-WTP method in health.

1.2 Economic evaluation in health using Quality Adjusted Life Years (QALYs)

Economic analysis in health care is conducted to aid in priority setting among competing health care programmes. Where the focus is the maximization of health gains, such as the UK, cost effectiveness analyses are preferred (National Institute for Health and Care Excellence (NICE) 2013; Weinstein & Stason 1977; Nimdet et al. 2015). Standard techniques are used to estimate the cost of programmes. The measure of effect used for the CEA is the quality adjusted life years (QALY). The QALY metric is used to assign values to disease burden (NICE 2013). This generic measure includes both the quality and quantity of life lived. One QALY equates to one year of life lived in perfect life. A threshold is determined as a cut-off for judging the cost-effectiveness of an intervention. The cost-effectiveness threshold is defined as the maximum value of money per health outcome that would be paid for adopting a given intervention or technology (Gold et al. 1996). In the UK, cost per QALY estimates are favoured and appraised in accordance with the £20,000 - £30,000 threshold (National Institute for Health and Care Excellence (NICE) 2013). An intervention is therefore adopted if the cost of a QALY does not exceed this threshold.

Other similar generic measures include the disability adjusted life years (DALYs) and healthy years equivalent (HYE). The DALY measures life years lost from disease, adjusted for assumptions about disability as well as the impact of age and future time (WHO 2012; Fox-Rushby 2002).Critics of the DALY approach argue that the measure:

(1) Is not inclusive of all population and disease groups;

(2) Offers a reductive view of health;

(3) Ignores the context within which diseases occur;

(4) Would increase inequalities of health across populations if minimisation was adopted;

(5) Does not reflect the preferences of the individual or the society in a given setting and therefore;

(6) Does not lead to the identification of the most efficient or welfare maximising interventions (AbouZahr 1999; Fox-Rushby 2002).

Similarly, the HYE combines quality and quantity of life into a single outcome metric for use in decision making (Mehrez & Gafni 1991). Unlike QALYs which only represent a patient’s part-preferences (the quality), HYEs capture full preferences based on the manner in which they are derived from the utility functions of each individual (Mehrez & Gafni 1991). However, some researchers in the field have argued that HYEs can only capture the full preferences if the utility function for reference flows of health is linear, which is not always the case (Ried 1998). Overall critique of the HYEs method however, relates to the complexities in application in the measurement of health outcomes (Wakker 1996; Hauber 2009).

While the QALY is recognised as a better measure of health outcomes or benefits, when compared to other generic measures, limitations inherent in the method have generated discussions on possible alternatives, including CBA. Primarily, the CEA culminates in the determination of a cost per additional QALY gained. While this ratio is useful for comparison among alternatives, it still does not provide information about whether the society believes that it is worth paying for the good or service or not and neither does the ratio translate to affordability (Cairns 2016). A key concern with the QALY measure is that it only accounts for only part of the preferences which are known or expected to underlie individual valuation of an intervention, goods or services (Liljas & Blumenschein 2000b; Labelle & Hurley 1992; Weatherly et al. 2009). As a result, estimates obtained in this way do not reflect the total value. Some of the other limitations with the use of QALYs as discussed in the literature are outlined below.

1. There is an assumption that a QALY is a QALY, i.e. the same QALY applies to all individuals when valuing any particular intervention. This assumption ignores the fact that health care gains (QALYs) differ by individuals (Edwards et al. 2013).

2. The above assumption also introduces equity concerns. In particular, the value of the sum of the health gains from any given intervention is expected to depend on the distribution. However, with the above assumption (a QALY is a QALY), estimates are based on the assumption that one QALY bears the same weight across different individuals regardless of their background characteristics including age, ill-health status and socio-economic characteristics (Olsen & Donaldson 1998). However, individual dynamics influence the outcome of interventions. For instance, patients would be expected to react differently to an intervention, e.g. pain relief medication, depending on the stage or severity of their disease. This is also likely to be influenced by the age. Studies have rendered support to discriminating in marginal QALY gains, for some of these attributes (Bleichrodt 1997; Olsen & Donaldson 1998).

3. The QALY approach relates primarily to health benefits. Valuation using the method assumes that non-health attributes can be ignored (Labelle & Hurley 1992). However, the consumption of health, leads to both health and non-health benefits (Olsen & Donaldson 1998). For instance, a pain management intervention may do that effectively. With the reduced or eliminated pain, an individual may be able to engage in activities which they may have previously paid others to do for them, such as household chores. Failure to account for benefits such as these leads to incorrect estimates of benefit.

4. The QALY measure also ignores the role of externalities in the generation of health. Good health does not occur in a vacuum and is often the result of a combination of inputs from multiple sources. For instance, an individual may be

supported to regain good health by a network of family and friends, the availability of a clean environment and facilities and other micro-level and macro-level factors. By relating the outcome of interventions to individuals only, an analysis using QALYs could potentially lead to non-optimal allocation of resources (Labelle & Hurley 1992; Weatherly et al. 2009). The effect of individual behaviour change beyond the length of interventions is also largely ignored with QALYs (Kelly et al. 2005).

The above limitations of the QALY measure are not exhaustive. There is consensus among researchers in the field on the need for a more comprehensive measure of outcome (Edwards 2001; Payne et al. 2013; Weatherly et al. 2009; Labelle & Hurley 1992; Olsen & Donaldson 1998). Among the suggested methods is the CBA approach, with WTP used to measure benefit.

1.3 The contingent valuation method as an alternative to the QALY

The CBA, using contingent valuation (CV) techniques is a particularly powerful tool where markets do not exist or would fail (for example due to the existence of public goods), or have become potentially politically excluded (such as the health sector) (Drèze & Stern 1987; McIntosh et al. 1999; McIntosh et al. 2010; Drummond et al. 1987; Mishan 2016). In such situations, it is difficult to determine a price for the good or commodity of interest, hence the use of survey techniques to estimate price. Compared to the other economic evaluation techniques, the CBA has a firm theoretical basis in economic welfare theory, as will be discussed in chapter 3.

Societal welfare is assumed to be maximised by undertaking only those interventions which have a positive difference of net present benefits minus net present costs (Drèze & Stern 1987; McIntosh et al. 1999; McIntosh et al. 2010; Drummond et al. 1987; Mishan 2016). By adopting a societal perspective, the CBA through CV-WTP studies, addresses some of the limitations inherent with the QALY approach including; valuation of non-health benefits, consideration of externalities, valuation beyond the individual to the society, accommodation of both use and non-use values and permits the assessment of inter-sectoral costs and consequences. Values obtained using the WTP technique can be used in a variety of ways including: decision making about investment in health care programmes; determination of the demand for private consumption goods (including some quasi-public goods), as a function of price; how much individuals are willing to pay for a service at the point of consumption and, setting insurance premiums (Gafni 1991; Olsen & Smith 2001; Olsen & Donaldson 1998; Bala et al. 1999).

1.4 Reported challenges with the use of the CV-WTP method in health

However, despite the apparent strengths of the method, the adoption of the CV-WTP technique in the health sector, compared to other sectors such as environment has been limited. One proponent of the method, Gafni (1997) interestingly observes that "*in the same period that the CV method has evolved and become the most commonly used method of valuing environmental benefits, the development in health economics has instead been towards CEA and CUA*". It is interesting to note too that the CBA evaluation technique is the most commonly used economic evaluation technique in most fields other than health care (Gafni 1997).

Among the challenges with the use of the CV method in health as suggested in the literature are the perceived difficulties in valuing health benefits (Bala et al. 1999). "Health", or improvements in individual or population health is the outcome of most interventions and this is primarily "intangible. In particular non-economists have been reluctant to place dollar values on the benefits of health care (Liljas & Blumenschein 2000b). This relates to both the conceptual challenges of valuing health, especially in settings where health care is publicly funded, to the methodological challenges relating to the valuation process. In all settings, health care is a scarce commodity, with limited resources available to meet all the needs. As discussed in an earlier section, decisions must be made about health care services and commodities to invest in, the level of investment and potential individual and population benefits of such investment. Willingness to pay values can therefore be used in setting priorities. Conceptually, in health care systems where patients do not have to directly purchase key components of their health care, making monetary valuations poses significant challenges (York 2016). However, even in such settings, with some guidance, health system users may able be to place a WTP value on specific components such as waiting times, types of care and where this is usually obtained through direct payments (*ibid*). For instance, evidence has shown that service users clearly perceive and can value (using WTP), the economic value of nursing services (Martín-Fernández et al. 2013). In this study, the authors used the CV method to value nursing consultation services from among users in a primary care setting.

Study respondents were able to value the services with the authors concluding that the CV method was useful for making explicit users' perception of value of health services (*ibid*). In the majority of settings where health care is not publicly funded, patients pay directly out of pocket to access health care services and commodities. This has led to arguments regarding financing of health care services and the elicitation of values from patients or population groups. One such argument that has been offered relates to the ethics of valuing health.

Secondly, some authors have argued that WTP estimates obtained from CV studies are not consistent with economic theory (Diamond & Hausman 1994). These authors suggest that, for instance, study respondents do not consider their income or budget constraints in valuations using the WTP approach. As a result, inconsistent estimates are obtained. The inconsistencies between income and stated WTP values have been established in studies across sectors (Hanemann 1991; Bateman et al. 1997; Sugden 1999). The preponderance of evidence from empirical assessments of the relationship between income and WTP values suggests a relationship between income and stated WTP. Specifically, in line with economic theory, respondents who are less wealthy are expected to have less disposable income and therefore state lower WTP values, *ceteris paribus*, compared to those from wealthy households. However, this relationship does not always consistently hold and WTP studies continue to investigate the effect of income on stated WTP. There are also concerns about how the distribution of income is incorporated with the CV methodology (Bala et al. 1999). Realistic statements regarding individuals' WTP must be bounded by an individual's ability to pay. In many contexts, the distribution of income is not expected to be equal across population groups. Authors argue that with unequal distributions of income, comparisons of the dollar values for various persons is difficult and often involves assumptions on equity that may be considered too strong (Diamond & Hausman 1994). Some studies conducted to establish the validity of the above mentioned claims, however, concluded that the critiques were based on poorly conducted CV studies (Portney 1994). Hanemann (1994), further established that some of the issues suggested in critique of CV-WTP studies were in fact, valid for all empirical studies (Hanemann 1994). Other factors such as the type of valuation good, WTP payment vehicle and WTP elicitation technique discussed in a later section, have been shown to have greater influence on stated WTP values.

The greatest critique of CV-WTP studies relates to uncertainties about the validity of WTP values (Liljas & Blumenschein 2000b; Telser et al. 2008; Loomis et al. 1996a; Loomis et al. 2006; Carson et al. 1996; Harrison & Rutström 2008; Little & Berrens 2003; Blumenschein et al. 1998). While some of these authors argue that the correlation of hypothetical values with actual behaviour is largely unknown most of critiques argue that hypothetical WTP significantly over-predicts actual values. This disparity between hypothetical and actual values is referred to as hypothetical bias and is a measure of the criterion or external validity of WTP. The concept of hypothetical bias is discussed in detail in subsequent chapters.

The debate on the use of CV-WTP has been long standing with authors such as Diamond, Hausmann and Portney (1994) calling for economic researchers to be involved in improving the method (Portney 1994). In the first large scale examination of the CV method, the NOAA panel acknowledged the strength of the method, while also noting potential biases inherent with the technique (Arrow et al. 1993). In their report, the NOAA panel offered some guidelines for addressing some of these biases. Since then, the method has been evaluated variedly in different sectors, with promising results. Methods for minimising the hypothetical bias have also been suggested and tested. For instance, in the first ever within-sample field test of the method, Ryan et al (2016) elicited WTP values for a health good using the bidding technique. The authors established the presence of hypothetical bias (hypothetical yes exceeded actual yes responses) in their results. However, calibration of responses minimised the discrepancies between hypothetical and actual responses. Authors established a constant rate of response reversals across the bids and concluded that this suggested theoretically consistent option values (Ryan et al. 2016).

Given the strengths of the method, as discussed above, and in light of the limitations of the current economic evaluation technique used in the health sector, the CV-WTP offers the greatest potential for maximising welfare gains. Improving the method calls for an investigation of the methodological issues in the conduct of CV studies. These could be addressed for more meaningful conclusions, especially on criterion validity.

The analyses and discussions presented in this thesis seek to fill some of the gaps in knowledge in this field as discussed in the next section.

1.5 Addressing the gaps: contribution of the thesis

This thesis contributes to this field by exploring the criterion validity of contingent valuation WTP studies, identified as a key concern and barrier to the use of CV-WTP values in the health sector. The thesis objective is not to assess the criterion validity of CV-WTP methods, but, to critique the methods used to assess the attribute. The contribution of the thesis is five-fold:

1. I critically appraise the literature on the methods used to assess the validity of CV-WTP studies. The systematic review in chapter 4 highlights the variety in among others, definitions of validity, validity assessment methods, study designs and the range of goods valued. Significantly, this review highlights the limited number of validity assessments in the health sector.

2. In the first section of chapter 5, I critically evaluate the methods that have been used to assess the criterion validity of CV-WTP method. In the second section of the chapter, I review the current evidence on the external validity of CV-WTP studies. Notably, the last systematic review on the subject was conducted more than one decade ago. This review highlights similar issues as the general validity review presented above. In addition, and of significant relevance to the thesis aims, the limited criterion validity assessments in the health sector are noted. Further, this systematic review establishes that conclusions on the criterion validity of CV-WTP are more diverse than authors are reporting in the primary studies.

3. Using a random effects meta-analysis, I quantify the magnitude of hypothetical bias from current external validity assessments. Again, the last quantitative summary was conducted more than one decade ago. Since then, several external validity assessments have been conducted, providing more estimates from a wider range of variables. These have been further explored and a range of potential drivers of hypothetical bias identified.

4. Using an empirical dataset, I illustrate the impact of elicitation techniques on hypothetical WTP values. Specifically, I analyse hypothetical WTP dataset elicited using a bidding technique, and open-ended methods from the same respondent. Predictions of actual WTP are made at the different bid levels and the open-ended

question. The drivers of WTP at the different bid levels and with the open-ended question are also determined. The results of these analyses demonstrate the effect that the ultimate choice of summary hypothetical WTP value significantly influences criterion (external) validity estimates.

5. In making the above contributions to the debate on the criterion/ external validity of CV-WTP studies, this thesis highlights some of the issues in the design and reporting of external validity assessments. A suggestion is also made for the development of guidelines for the conduct and reporting of criterion/ external validity assessments. Based on the results of the various analyses presented in this thesis, I have provided suggestions of what these guidelines might look like, in early attempts to improve the method.

In the next chapter, the framework for the analyses presented in this thesis is presented.

Chapter 2 Conceptual Framework

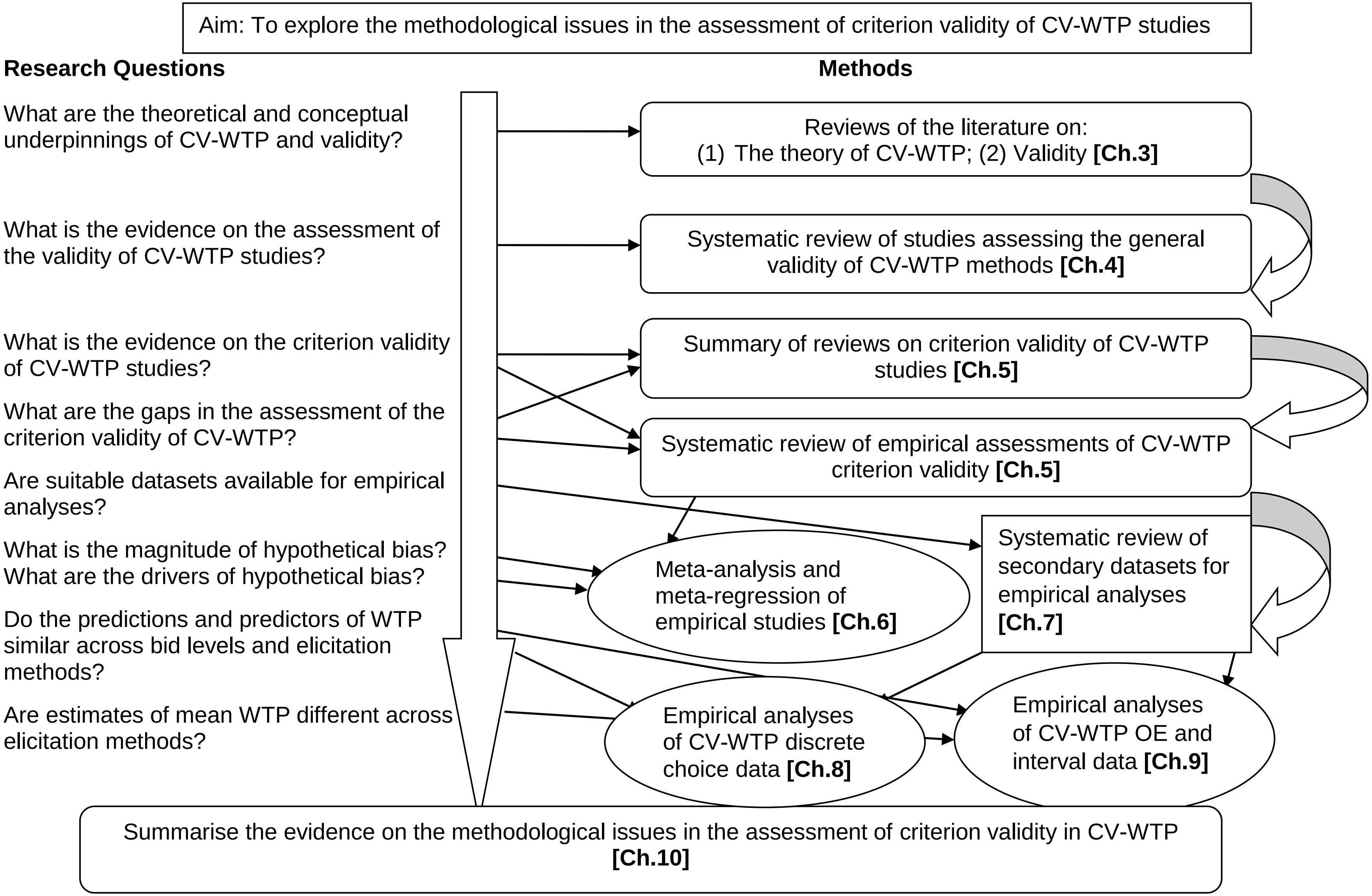
2.1 Introduction

As was discussed in the last chapter, the criterion validity of CV-WTP studies is the subject of ongoing criticism. This thesis aims to contribute to knowledge in this field by highlighting some of the methodological issues in criterion validity assessments. A mixed methods approach, covering literature reviews and empirical analyses was used to address the aims of the thesis. In this chapter, these methods are specified. The chapter is structured as follows: In the next section, the conceptual framework for the thesis is outlined (figure 2.1), clearly illustrating the research questions, related activities and the links between the chapters. The methods used to address the thesis objectives discussed in section 2.3. Finally, the layout of the thesis is presented with a description of the respective chapters. This chapter lays the foundation for the thesis and reference will be made to the methods outlined here throughout the document.

2.2 Conceptual framework

To address the objectives of this thesis, a series of research questions were developed. Each research question was answered using a combination of distinct but related methods, as illustrated in figure 2.1. The chapters linked to each of the research questions are also indicated.

Figure 2-1: A schematic overview of the conceptual framework


Aim: To explore the methodological issues in the assessment of criterion validity of CV-WTP studies
Research Questions
Methods
What are the theoretical and conceptual underpinnings of CV-WTP and validity?
What is the evidence on the assessment of the validity of CV-WTP studies?
What is the evidence on the criterion validity of CV-WTP studies?
What are the gaps in the assessment of the criterion validity of CV-WTP?
Are suitable datasets available for empirical analyses?
What is the magnitude of hypothetical bias? What are the drivers of hypothetical bias?
Do the predictions and predictors of WTP similar across bid levels and elicitation methods?
Are estimates of mean WTP different across elicitation methods?
Reviews of the literature on: (1) The theory of CV-WTP; (2) Validity [Ch.3]
Systematic review of studies assessing the general validity of CV-WTP methods [Ch.4]
Summary of reviews on criterion validity of CV-WTP studies [Ch.5]
Systematic review of empirical assessments of CV-WTP criterion validity [Ch.5]
Meta-analysis and meta-regression of empirical studies [Ch.6]
Systematic review of secondary datasets for empirical analyses [Ch.7]
Empirical analyses of CV-WTP discrete choice data [Ch.8]
Empirical analyses of CV-WTP OE and interval data [Ch.9]
Summarise the evidence on the methodological issues in the assessment of criterion validity in CV-WTP [Ch.10]

2.3 Methodological approaches

To address the thesis aims, a four-pronged approach was adopted:

1. Reviews of the theoretical and conceptual underpinnings of the thesis subjects. The theoretical reviews provide the philosophical boundaries within which the work presented in the thesis is situated. Two reviews were conducted:
   i. Theoretical underpinnings of the contingent valuation methods;
   ii. The historical perspectives on validity.
2. Systematic reviews to establish the empirical evidence on validity. The purpose of this was to investigate what has already been done on validity of CV-WTP methods, evaluate how this has been done and establish the gaps in knowledge. In addition, a review was conducted to determine predictors of WTP for malaria nets. These informed the empirical analyses presented in the next section. The following four systematic reviews were conducted:
   i. A systematic review on the assessment of the main types of validity of CV-WTP methods. The purpose of this was to critically appraise the methods that have been used in the assessment of CV-WTP validity and identify gaps in knowledge. Some of the identified gaps are addressed in subsequent chapters.
   ii. A focussed summary of reviews on the criterion validity of CV-WTP studies. The purpose of this was to evaluate the evidence on the criterion validity of CV-WTP and the methods that have been used in such assessments. The methods were used to inform the design of the subsequent review of empirical studies assessing the criterion validity of CV-WTP. The review was also used to establish the evidence on the variables which influence hypothetical bias in CV-WTP. These independent variables were further explored in the first empirical analysis quantifying the magnitude of hypothetical bias in primary studies assessing criterion validity.
   iii. A systematic review on the criterion validity of CV-WTP methods. Based on the findings from the previous summary of criterion validity assessments, the purpose of this review was to critically analyse the empirical evidence on hypothetical bias, the methods that have been used to assess this and demonstrate the gaps in knowledge in this field. The

findings from this review further informed the empirical analysis discussed in a later section.

iv. A review of studies investigating willingness to pay for malaria treated mosquito nets (TMN). The purpose of this review was to explore the independent variables that have been investigated in the assessment of WTP for TMN. These variables informed the construction of models tested in two of the empirical analyses in this thesis.

3. A systematic review of potential datasets for use in addressing this aims of this thesis. To address the aims of this dissertation, an empirical dataset was needed. A primary study could have been designed and appropriately targeted to serve this purpose. However, as the focus of this thesis is purely methodological, re-invention of this wheel was deemed unnecessary. Further, the use of secondary data has been lauded as a cost-effective way of making full use of primary data that are already collected (Cheng & Phillips 2014; Vartanian 2011). The dataset identified in this stage was used for the empirical analyses discussed below and presented in chapters 8 and 9.
4. Empirical analyses. Given the gaps identified in the literature, the following three distinct but related empirical analyses were conducted in this thesis.

   i. A quantification of the extent of the magnitude of hypothetical bias from systematically reviewed studies. The drivers of hypothetical bias were also explored.

   ii. The analysis of discrete choice CV-WTP data. Univariate and multivariate analyses were conducted to illustrate the differences in both the predictions and the predictors of WTP at different bid levels and with different methods. The analysis also demonstrates the potential impact of these differences on criterion validity conclusions.

   iii. The analysis of open ended and interval CV-WTP data. This analysis was conducted to further illustrate the multiple estimates of WTP that can be obtained with different WTP elicitation techniques even on the same population group. The impact of the different estimates on criterion validity conclusions is further demonstrated.

2.4 Organization of the thesis

This thesis is composed of ten chapters, including the introductory and the current chapter. The rest of the chapters are organised as follows.

The theoretical foundations of the contingent valuation method are evaluated and summarised in chapter 3. In this chapter, the design and analysis of contingent valuation-WTP studies is discussed. The use of the CV-WTP method is the subject of ongoing criticism, with the validity of the method questioned. Concerns relate to whether values revealed using hypothetical surveys of WTP methods correctly predict expected actual values. The concept of validity as applied in economic evaluations is introduced and the different types of validity outlined. This thesis explores some of the methodological issues in the design of CV-WTP studies which may contribute to hypothetical bias, and hence conclusions on criterion validity. This chapter therefore lays the theoretical framework for subsequent discussions in the thesis.

In chapter 4, the literature on the assessments of the validity of WTP methods in health is critically appraised using a systematic review. This review highlights the variety in validity terms and definitions and the assessment methods. While criterion validity assessments have been recognised as the most definitive tests of the validity of contingent valuation WTP methods, relatively few such studies are identified in the literature. In addition to the variety in the study designs, there does not seem to be any consistency in the manner in which studies are reported, data analysed, or the conclusions made thereof, regardless of the type of validity assessed. In this chapter, the scarcity in empirical assessments, and inconsistencies in the methods used in the assessment of criterion validity is demonstrated.

In chapter 5, two reviews are presented. The first is an evaluation of the methods used to assess the criterion validity of WTP methods. Based on the limited number of empirical studies assessing criterion validity identified in the systematic review of all types of validity presented in chapter 4, the purpose of this review was to identify previous reviews of criterion validity. To assess criterion validity, values obtained from contingent valuation surveys have been compared with varied techniques presented in figure 1.1. This scoping review provides a justification for the focussed systematic review on the criterion validity of WTP methods presented. These reviews are also used to inform the design and methods for the second review presented in this chapter. This is a focussed critical appraisal of empirical assessments of criterion validity. The last systematic review on the subject was conducted more than a decade ago. The systematic review summarises the evidence on hypothetical bias in CV-WTP studies conducted in across sectors, highlighting the gaps in among others, empirical assessments in health and lack of standardised reporting guidelines for the studies. Interestingly, also, most of CV-WTP studies were conducted in low and middle-income countries, and in countries where health care is not publicly financed. Studies were conducted more for the purposes of testing hypothesis and improving the method, but not for decision making as is the case in sectors such as environment. Variety is noted in the methods used in the criterion validity assessments, and the reporting of key study results and attributes. Further, the conclusions on the presence of hypothetical bias are mixed.

In chapter 6, I quantify the magnitude of hypothetical bias (the disparity between values obtained from the hypothetical survey and SMEs) from the reviewed studies. Based on the analysis from this larger dataset, the magnitude of hypothetical bias is lower than was established in the last meta-analysis, and therefore smaller than has been proposed by critiques of the method. The findings of the earlier systematic review and meta-analysis suggest that some methodological issues may lead to inaccurate estimates of WTP values, and hence incorrect assessments and conclusions on criterion validity. A meta-regression identifies some potential drivers of hypothetical bias and these are discussed. The chapter also highlights the need for guidelines for the conduct and reporting of criterion validity assessments in CV-WTP studies.

Chapter 7 discusses the process used to identify a relevant empirical dataset for analysis to address the thesis aims. Following the pre-determined criteria, ten potential datasets are evaluated and one of these selected. The dataset is discussed in this chapter. The data is based on an economic evaluation conducted in Surat, India. As part of a large randomized control trial examining the cost and effectiveness of different mosquito prevention interventions, a contingent valuation study was designed. In the hypothetical survey, the study assessed the willingness to pay for insecticide treated mosquito nets, among other interventions. The study used bidding techniques followed by an open-ended question to elicit WTP. This provided for the estimation for WTP at different levels, and the related bid functions. The study also collected data on a range of variables which allows for the testing of different drivers of hypothetical WTP and which might influence actual values. The limitations of the dataset are presented in concluding the chapter.

In chapter 8, an analysis of discrete choice responses is presented. The analysis is based on the data discussed in chapter 7. Discrete choice responses from the two bid levels and single dichotomous choice question are analysed. The aim of this is to illustrate the differences in both the predictors and predictions of WTP at the different bid levels. The impact of these on criterion validity assessments and conclusions is demonstrated.

Chapter 9 presents an analysis of open ended and interval data from the same dataset. The purpose of this analysis is to demonstrate the multiple estimates of mean WTP that can be obtained with the use of multiple elicitation formats. The predictors of WTP with the open-ended data are also determined. The results are used to demonstrate the effect of the elicitation and analytical methods of hypothetical WTP mean estimates, and the effect of these on criterion validity conclusions.

Chapter 10 summarises the thesis thread, highlighting the implications of the different reviews and analysis for policy and research. Based on these findings, I argue for the revisiting of the discussion on the contingent valuation-willingness to pay method as a powerful benefit assessment tool. I propose the development of specific guidelines for the conduct and reporting of criterion validity assessments. This is informed by current evidence on the disparities in the methods used in the conduct and analysis of criterion validity assessments. Based on the analysis presented in the thesis, I provide some suggestions for what these guidelines might include. In the conclusion I also argue for the conduct of more CV studies using the WTP method in the health sector where the method is still scarcely used.

The outputs in chapters 5 and 6 have been presented in national and international health economics meetings. Manuscripts developed from the different dissertation chapters will be submitted to peer reviewed journals. Throughout the thesis, efforts have been made to keep key outputs within the chapters for clarity, with further analyses presented in the appendix and referenced clearly within the chapters.

Chapter 3 Theoretical Framework of Contingent Valuation Methods

3.1 Introduction

The aim of this chapter is to introduce contingent valuation methods and discuss their use in benefit assessment for non-marketed goods. The theoretical framework underlying the contingent valuation techniques is provided in section 3.2. This is followed by a discussion focussing on the use of contingent valuation methods for benefit valuation. Key design attributes of a contingent valuation – willingness to pay study are outlined and the analysis of CV-WTP data discussed. The concept of validity is introduced, and the historical perspectives presented in section 3.3. The various forms of validity as applied in CV-WTP studies are discussed in section 3.4 and the chapter is summarised in section 3.5. Broadly, this chapter lays the theoretical foundation for subsequent discussions in the thesis and reference will be made to the concepts outlined here throughout the thesis.

3.2 The Contingent valuation method

3.2.1 Historical perspectives

The contingent valuation method has its origins in the environmental and natural resource sectors and has been used primarily to elicit preferences for public and quasi-public goods. The earliest use of the method is traced to the valuation of outdoor recreation in a Mane backwoods area (Davis 1963b; Davis 1963a; Davis 1964). Following on from this seminal work, CV valuations have been conducted to measure the benefits of different goods including recreation (Walsh et al. 1983), hunting (Cocheba & Langford 1978), water quality (Gramlich & Rubinfeld 1983), decreased mortality risk from a nuclear power plant accident (Mulligan 1978), and toxic waste dumps (Smith et al. 1985). More recently, the CV method has been used in the valuation of goods in other sectors including health and environment (Onwujekwe et al. 2002; Onwujekwe et al. 2001; Bhatia & Fox-Rushby 2002; Loomis et al. 1996; Loomis et al. 2006; Alberini 1995; Ahlheim et al. 2010).

In the history and development of the CV method, Michell and Carson (1989) single out the study by Randall, Ives and Eastman (1974) as the most influential. This study was notable for: (i) its theoretical rigour, (ii) the valuation of a good which could not be valued by any alternative method, (iii) the use of photographs in the hypothetical scenario setting to depict the visibility levels being valued (and therefore enhance the understanding of the respondent), and; (iv) the experimental design which involved the systematic variation of certain aspects of the bidding game (such as the payment vehicle) to determine whether the WTP amounts would be affected in a systematic manner (Mitchell & Carson 1989). Further credence to the CV method was given by the NOAA panel in their ruling and guidelines (1993), recognising it as a credible benefit valuation method and also providing some initial recommendations on the design (Arrow et al. 1993). The theoretical rigor in the design and implementation of CV studies is still the subject of much debate among economists and will be discussed throughout this thesis.

3.2.2 Theoretical Foundations

The CV method has its foundations in applied welfare economic theory, based on normative economics. Unlike positive economics which describe how the world works, normative economics describe how the world could work, leaning on the desirability of governments to undertake certain policies (Mitchell & Carson 1989; Carson 2000; Bateman & Turner 1993). The CV method aligns towards the Pareto criterion, which states that policy decisions which make one person better off without making anyone worse off should be adopted. As discussed earlier, the purpose of conducting a CV survey is to obtain an accurate estimation of the benefits (and sometimes costs) of a change in the level of provision of a good or service (Bateman & Turner 1993; Bateman et al. 2002; Mitchell & Carson 1989; Carson 2000; Slothuus et al. 2002; Bayoumi 2004).

The benefit – cost analysis (or cost-benefit analysis (CBA)) such as this is a variant of applied modern welfare economics. The CBA operationalises a variant of the Pareto criterion by using the CV techniques (WTP/WTA) to place a monetary value on the gains or losses expected from the change in the provision of a good or service. In this way, the net loss or gain can be calculated based on which the Pareto-efficiency of the change is determined (Mitchell & Carson 1989). In ensuring Pareto-efficiency, the assumptions of consumer sovereignty[6] and economic efficiency, based on positive economics, are adopted. While economic efficiency is a primary emphasis, of key concern in the analysis of economic data is the distributional outcomes of benefit valuation. The CV method is very unique among

[6] Consumer sovereignty: The belief that a consumer is the best judge of their own utility (Hutt, 1940)

the benefit measurement techniques in its ability to obtain detailed distributional information (Bateman & Turner 1993; Bateman et al. 2002; Carson 2000; Mitchell & Carson 1989).

3.2.3 Benefit Measures

Both compensating[7] (CV) and equivalent[8] variation (EV) measures of welfare can be elicited using the CV method. These two measures of welfare stem from the consumer surplus concept. Consumer surplus is measured using the ordinary demand curve, also known as the Marshallian demand curve which holds income constant, while varying utility and price in the estimation. Problems with the use of consumer surplus, based on the ordinary demand curve, as a measure of benefit have been documented (Samuelson 1947; Silberberg 1978). In attempts to address these, (Hicks 1941; Hicks 1943; Hicks 1956) suggested an extension of the Marshallian demand curve, the Hicksian demand curve.

The Hicksian demand curve holds constant the utility level at the initial level (compensating variation or surplus) and at the specified alternative level (following the change in provision), known as the equivalent variation (or surplus). A payment or compensation is involved in the four measures and this is determined by the consumer's property right position regarding the valuation good. Property rights (whether the consumer has the right to sell the valuation or not) also determine whether willingness to pay (WTP) or willingness to accept (WTA) is elicited. Table 3-1 illustrates eight possible welfare measures based on the Hicksian measures.

Table 3-1: Hicksian welfare measures for contingent valuation surveys

Proposed Change	WTP	WTA
Quantity Increase	CS	ES
Price Decrease	CS; CV	ES; EV
Quantity Decrease	ES	CS
Price Increase	ES; EV	CS; CV

WTP: Willingness to Pay; WTA: Willingness to Accept; CV: Compensating Variation CS: Compensating Surplus; ES: Equivalent Surplus; EV: Equivalent Variation

[7] Compensating variation (CV) is a measure of the amount of money which must be taken or given to a consumer to cancel out the welfare gain or loss resulting from a price change after it occurs, bringing the consumer back to the original utility level (Varian, 2014).

[8] Equivalent variation (EV) is the amount of money which would have to be paid to the consumer to enable them attain the utility level possible with the new prices and income while facing base prices and only having base income (Varian, 2014).

3.2.4 Elicitation perspectives

When using the previously discussed Hicksian measures, we assume that consumers are certain about the amount of utility they will obtain from the consumption of the valuation good (Bateman et al. 2002, Hicks 1956). However, in practical terms, it is not always possible to determine the utility that would be derived from the consumption of a good before it is offered. Further, the utility that one expects to obtain before knowing what having the good is like may differ from the utility the consumer expects to receive after experiencing the good. The former state is referred to as an “ex-ante” perspective while the latter is referred to as “ex-post”. The ex-ante perspective has been suggested as the ideal in most welfare economic analyses involving uncertain outcomes (Graham 1981; Chavas et al. 1986).

3.2.5 The design and conduct of a CV willingness to pay study

A CV study primarily consists of four key components: (i) the valuation good and the hypothetical scenario or market in which the good is made available to the respondent for valuation; (ii) the elicitation question; (iii) a series of questions which detail the respondent’s characteristics and (iv) the administration method. These components of a CV study are discussed in the next section.

3.2.5.1 The valuation good and the hypothetical scenario

In contingent valuation, the valuation of a good is conducted contingent on a market existing (McClelland et al. 1992). Therefore, a hypothetical scenario is designed which details the purpose of the valuation and the hypothetical market. The researcher constructs this model hypothetical market and in this discusses the hypothetical nature in which the good is to be valued. The hypothetical market is defined to be as plausible as possible, depending on the nature of good. A hypothetical scenario which is believable passes the face validity criteria which will be discussed in a later section. It provides sufficient detail about the valuation good, the baseline provision of the good (where applicable), the structure under which the good is to be provided, the range of available substitutes (if any), and the method of payment (Mitchell & Carson 1989). Hypothetical WTP values refer to the stated estimates that respondents provide when the hypothetical market is described. In this scenario, they are not expected to make a payment or take ownership of the valuation good or service.

3.2.5.2. The elicitation question or method

This refers to the type of question used to elicit values and is a key attribute of the CV method. Elicitation questions are designed to facilitate the valuation process without themselves biasing the respondent's WTP (Carson & Louviere 2011; Bateman et al. 2002; Mitchell & Carson 1989). The different WTP elicitation questions are presented in Appendix 2. The choice of the elicitation question depends primarily on the nature of the valuation good and the study administration method. Some elicitation methods are also associated with higher response rates, for example the dichotomous choice question. This is because, unlike other methods, this method closely mimics a market transaction which most respondents will be already familiar with. As a result, the cognitive challenge with this method is lower than is the case with other methods. However, the DC method does not permit the measurement of the respondents' maximum welfare (consumer surplus). On the other hand, the open-ended method allows for the estimation of the respondent's consumer surplus. However, the method is associated with higher outliers and zero values.

The type of WTP data obtained is determined by the elicitation question. Based on the question, WTP data can either be discrete or continuous. Continuous data is obtained with open ended questions while the other question types elicit discrete choice WTP data. Associated with the elicitation question also is the payment vehicle and frequency of payment. The payment vehicle refers to the method which is used to collect the money. Payment vehicles include taxes on property or services, donations or voluntary contributions to a cause and out of pocket payments. Payments for valuation goods can be collected as a one-off, on a monthly or annual basis or at the point of sale for a given commodity.

3.2.5.3 Background questions on the respondent

Respondents' demographic and socio-economic characteristics such as age, sex, education, occupation and a measure of wealth such as income or assets are also documented. Additional information may be collected depending on the valuation good. These may include experience with a commodity or service or related commodities / goods, preferences for a given good / service and the reasons for the valuation choices given. This information can be obtained before or after the hypothetical scenario has been set. These set of questions are used to estimate a valuation function for the good and explain the WTP values obtained in the study (Arrow et al., 1993; Bateman et al. 2002). When the estimation obtained using variables identified in theory as predictive of WTP is positive, this is taken as some evidence for reliability and validity of the study and / or estimates (Mitchell & Carson 1989).

3.2.5.4 Study administration mode

Contingent valuation studies can be administered through mail, telephone or face-to-face interviews. Administration modes are informed primarily by the elicitation question. These are also determined by other factors such as the resources available (time and financial), quantity of data and data control expected, required response rate, the degree of complexity of the valuation (and good) and the versatility allowed (Bateman et al. 2002). Face-to-face interviews are considered to be the most resource intensive. Despite this, they are also associated with the highest response rates. Face-to-face interviews have been recommended by the NOAA panel (Arrow et al. 1993) for use in the assessment of natural resource benefits. The effect of interviewer bias in face-to-face interviews (which is lowest with mail interviews) has been observed but some elements of this can be corrected through careful and rigorous training of interviewees prior to the start of the survey (Bateman et al. 2002). Mail surveys are the cheapest, but they are also associated with very low response rates. Telephone surveys, considered an in-between of mail and face-to-face interviews, are faced with low response rates with respondents unwilling to participate in cold-call interviews that last for more than a few minutes, unless they are truly interested in the survey.

The theoretical underpinnings of the CV method have been discussed in the above section. In addition, the design and administration attributes as outlined in theory have been summarised. In the next section, biases inherent in WTP studies are presented.

3.2.6 Biases inherent in the analysis of WTP data

In the analysis of both open ended and closed ended data, respondents express their unwillingness to pay for the valuation good in different ways. In the open-ended data, this could be by providing values that are too low, including zero values, or too high (and often not correlated with other respondent variables such as income). In the closed ended data respondents are expected to provide a “Yes” or “No” response. A “No” response in this data could be interpreted as either a true zero value placed on the valuation good or that a respondent is not willing to pay for the good. For both open ended and closed ended data, the correct interpretation of zero or “No” responses is related to the accurate estimation of WTP values. A follow up question is often asked to help in the accurate classification of the “No” responses.

In the absence of a follow up question, one could assume that the zero (with open ended data) and “No” (with closed ended data) represent true zero values, meaning that their WTP was zero, but these could also represent one of three broad classes of biases (Halstead et al. 1992). These are:

(1) Instrument related biases[9], strategic bias[10] or scenario mis-specification[11] (Mitchell & Carson 1989; Cummings Brookshire, D.S. and Schulze, W.D. 1986; Calia & Strazzera 2001);

(2) Protest bias where respondents choose not to respond to valuation questions, or place a zero value even when they have a positive valuation of the good or provide extreme values which can be regarded as outliers (Halstead et al. 1992);

(3) Whole or part non-response to the study questionnaire (e.g. failing to respond to the valuation question partly or in totality in a survey) (Dalecki et al. 1988; Halstead et al. 1992).

In the next section, starting point bias and protest responses which will be explored in this thesis are discussed further. Whole or part non-response was not encountered in the data used for the empirical analyses in this thesis and is therefore not explored further.

Starting point bias

Starting point bias occurs when the elicitation method or payment vehicle directly or indirectly introduces a potential value cue which influences the WTP amounts given by a respondent (Mitchell & Carson 1989). A major limitation of the bidding method

[9] Instrument related biases are caused by or occur in relation to the WTP instrument itself and include starting point bias.

[10] Strategic bias: Respondents respond strategically to the question if they believe that their answer might have some influence e.g. on the pricing of the valuation good or taxation.

[11] Scenario mis-specification occurs when the hypothetical scenario description is misunderstood by the respondent.

as evidenced in the CV literature is anchoring or starting point bias (Mitchell & Carson 1989; Cummings et al. 1986; Veronesi et al. 2011; Bhatia 2005; Roberts et al. 1985; Welle 1985). For the majority of the respondents, the starting bid offered implies a value of the good and therefore their responses are anchored around this first bid. This leads to a tendency for yea-saying (Mitchell & Carson 1989; Arndt & Crane 1975; Couch & Keniston 1960). Starting point bias has been observed in many studies, particularly in the environmental literature (Herriges & Shogren 1996; Flachaire & Hollard 2007; Chien et al. 2005; Alberini et al. 2005; Holmes & Kramer 1995).

Researchers in the field suggest that starting point bias explains the internal inconsistency with values obtained at different bid levels, where multiple bids are used to elicit WTP values. When a respondent anchors their stated WTP to an earlier bid, the estimated mean and the related dispersion of the estimate can be significantly biased. Different econometric models can be used to minimise or account for the effect of starting point biases. The effect of these biases can also be investigated by including the different bids as regressants in regression models. The relationship between the starting point bids and the WTP amounts is thus determined. However, other researchers have argued that the effect of ignoring starting point bias is complex and depends on the true distribution of WTP (Alberini et al. 2005).

Protest responses

While some respondents indicate “No” or provide zero values as their honest and correct estimation of the valuation good, others do so in protest to the whole idea of paying for the valuation good. When the latter is the case, this is referred to as protest responses (Halstead et al. 1992). Protest responses may also be expressed by respondents choosing not to participate in the valuation exercise. For example, respondents may indicate that they are not in the market for a commodity in the hypothetical survey but purchase the commodity in the actual survey. Failure to account for protest responses leads to incorrect estimations of aggregate WTP values.

Protest responses particularly pose a challenge in the analysis of discrete choice data as they are more difficult to interpret, particularly without a follow up question (Halstead et al. 1992). The challenge lies in the fact that by responding “No” to the bid question, it is possible that the respondents' willingness to pay is below this bid and not necessarily zero. In this situation the protest bid is therefore misinterpreted. A double bounded DC bid can help to better estimate the respondent's valuation in such a scenario. In this, respondents are offered additional option(s), expanding their range of choice in responses. To consider and, where necessary correct for the possible bias due to protest responses, sample selection models are specified. These are detailed in the discussion on the analysis methods in the next section.

3.2.7 The analysis of contingent valuation data

In a carefully designed, pretested and implemented study, respondents' answers to valuation responses can be taken as valid WTP responses (Mitchell & Carson 1989). However, the appropriate analysis of WTP data directly determines the validity of the benefit estimates thus obtained. This includes a consideration of the possible biases as discussed in the previous section, in the analysis.

The way respondents make and express their preferences in a contingent valuation setting is complex. This complexity can be increased or reduced by the choice of the WTP elicitation method. For researchers, the choice between the different elicitation methods, depends on the nature of the good, the purpose of the valuation process, the availability of resources (time and finances) to conduct the research, and the level of information required (Bateman et al, 2002). Some of the design attributes discussed in section 3.2.5 are also considered. The planned analysis of such data is also expected to influence the choice of elicitation method. The WTP analysis methods are categorized along two broad dimensions related to the question (Carson & Louviere 2011; Mitchell & Carson 1989; Carson 2000; Bateman et al. 2002). These are: (i) whether the actual maximum WTP (consumer surplus) for the good in question is obtained and; (ii) whether a single WTP question or an iterated series of questions is asked for the valuation good. This follows from the classification of elicitation methods as either open or closed ended (discrete choice), respectively.

The correct choice and application of analytical methods ensures accurate WTP estimates. However, as will be discussed in a later chapter, the methods used in the analysis of WTP values are rarely reported and /or, incorrect methods are used, questioning the validity of the estimates thus obtained. Even with the application of theoretically and statistically correct designs and methods, the validity of CV-WTP estimates has been questioned. The reported lack of validity has been the major reason cited for the limited use of the method, as compared to other methods, particularly in the health sector. In the next section, the concept of validity is discussed. In addition to the definitions, the different types of validity as used with contingent valuation studies are discussed. As this thesis focusses on the assessment of criterion validity, further considerations on this are provided.

3.3 The concept of Validity

Validity is defined in measurement theory as: a measure of the extent to which an instrument measures what it is expected to measure; a measure of the extent to which a concept is actually represented by the indicators of such concepts or a measure of the agreement between a test score or measure and the quality it is believed to measure (Yue 2010; Carmines & Zeller 1979; Parker 1990; Kaplan & Saccuzzo 1997). The main types of validity discussed in literature include face, content and criterion validity. The definitions of the different types of validity as adopted in this thesis are presented in Appendix 3. In the next section, historical perspectives on validity as applied to CV are discussed.

3.3.1 Historical perspectives on validity

Historically, validity was defined in terms of the accuracy of an estimate (Kane 2001). A variable of interest was assumed to have a definite value for each person, and the goal of measurement was to estimate this variable's value as accurately as possible. In psychology and education, a criterion measure was required to provide the 'real' value of the variable of interest or at least a better approximation of this 'real' value (Thorndike 1913). With the criterion measure, validity was defined as an assessment of how well the scores on a test agreed with what they were meant to measure (Kane 2001). This is what is currently regarded to as criterion validity. This type of validity required that suitable criterion on which to value any other measure existed. However, validation of criterion would be a cyclic, never ending process, leading to an infinite regression of criterion validation studies (Kane 2001). Further, well-defined and demonstrably valid criterion measures are not readily available and when they are available, they are subject to the same infinite circularity in comparing one criterion against another (Kane 2001).

Fields such as biochemistry resorted to the examination of the content of the items of the test to ensure that they adequately sampled the subject being evaluated. This is referred to as content validity. The underlying assumption in content validity was that if all aspects of the content were included, and there were no items that were irrelevant, then the test would be intrinsically valid in that it assessed whether or not the person had mastered the course content (Streiner & Norman 2008). Critics argued that content validity was highly subjective and tended to have a strong confirmatory bias (Guion 1974). The judgements about what a test item measures or the content domain covered by a test are usually made during test development or soon after, by persons involved in test development and such persons always see the test as a reasonable way to measure the attribute of interest (Kane 2001).

Further, Messick (1989) argued that content validity played a limited role in validation because it does not provide direct evidence even though it provided support for the ‘domain relevance and representativeness of the test instrument’ (Messick 1993). Critiquing both content and criterion validity, Messick (1993) further argued that in some domains such as clinical psychology, there is no objective criterion against which scales can be validated and content validity is insufficient because it does not provide evidence in support of inferences to be made from test scores (Messick 1993). It was not possible then to assess the usefulness of scales that were more widely used in this domain to evaluate areas such as attitudes, beliefs and feelings and pathological states like depression, anxiety and schizophrenia.

Construct validity was introduced in attempts to broaden the then definitions of validity to accommodate the interpretations assigned to clinical assessments (Cronbach & Meehl 1955). It is a framework of hypothesis testing based on the knowledge of the underlying construct such as anxiety or depression. The validity of the proposed interpretation of scores in terms of the construct is evaluated in terms of how well the scores satisfy the theory (Kane 2001). If the underlying theory is correct and if the test is valid, then the study should come out in the way that was predicted. With the unlimited hypothesis that can be made from theory, construct validity is a continual, never ending task of determining how a scale performs in a variety of situations (Streiner et al. 2008). With the introduction of construct validity, measurement theory adopted the ‘Trinitarian’ view of validity, dividing it into:

‘content’, ‘criterion’ and ‘construct’ validity; with construct validity further subdivided into convergent, discriminant and trait validity (Streiner et al. 2008).

A less discussed form of validity in the literature is face validity. This is the most basic form of validity and is concerned with how well a measure represents an intuitive and common-sense understanding of a phenomenon or how well a test or the questions on a test appear to measure the desired qualities of a particular construct (Bowling 2002; Carmines & Zeller 1979; Yue 2010; Streiner & Norman 2008).

3.3.2 Summary on validity

Until the 1960’s, validity was seen as demonstrating the psychometric properties of a scale and was often defined in terms of whether a test was measuring what it was thought to be measuring (Kelley 1927). Emphasis then changed to the characteristics of the people who were being assessed and the scores they achieved (Cronbach 1971). Validation processes would then be directed toward the inferences that can be made about the attributes of people who have produced those test scores (Landy 1986). Validating a scale is therefore a process through which we determine the degree of confidence we can place on the inferences we make about people based on their scores from a given scale (Streiner et al. 2008). Further, construct validity was reconceptualised to encompass all forms of validity testing – “all measurement should be construct-referenced” (Messick 1986).

Scholars in the field argue that construct interpretation is the basis of all score-based inferences. It is thus not just related to interpretive meaningfulness but also the content and criterion-related inferences specific to applied decisions and actions based on test scores (Messick 1986). Hence, discussions on validity focus primarily on the logic and methodology of hypothesis testing. Scholars such as Streiner and Norman (2008), view validity as a unitary construct but refer to different types of validity testing. Validation is thus regarded as a process with validity as the outcome (Streiner et al. 2008). The criterion validity of CV-WTP, the focus of this thesis, is the subject of ongoing criticism. However, while critiques focus on the summary estimates, questions can be asked of the methods used to derive these. Discussions also revolve around the reliability and scope (in)sensitivity of CV-WTP methods. This is briefly discussed in the next session.

3.3.3 Reliability and scope insensitivity in CV-WTP data

The reliability and scope sensitivity of WTP estimates has also been tested as a measure of validity. Reliability is defined in the literature as a measure of the extent to which an instrument measures intended attributes in a reproducible and consistent fashion (Streiner 1989). Bowling (2001) defines this concept as "the ability to produce consistent results, and consistent results on different occasions, when there is no evidence of change". Three forms of reliability tests are often discussed: (1) test-retest reliability in which a tool or scale is administered to the same population at different time points (Cronbach & Meehl 1955); (2) internal consistency in which a concept is tested using different scale items and; (3) intra-rater reliability which focuses on the researcher or interviewers (tests the consistency of a measurement by different interviewers or an individual interviewer at different time points) (Bowling 2001).

In WTP, the question of reliability focusses on the ability of respondents to reliably understand and answer WTP questions, given the hypothetical nature of the surveys. The cognitive burden of WTP questions given respondents from varied literacy levels is acknowledged (Foreit & Foreit 2003). However, the evidence on the effect of education or literacy levels on the ability to reliably respond to WTP questions is mixed. Lack of understanding has been shown to lead to two key reliability problems: (1) Non-response where respondents simply fail to respond to the valuation question or provide a don't know answer; (2) Yea-saying where respondents answer in the affirmative to closed ended questions while masking their true WTP value. Both issues have been discussed earlier in this section. The reliability of WTP values is tested based on the magnitude of the explanatory power of the regression model used in the study (Mitchell & Carson 1989). These authors propose an R^2 value of (R^2>0.15) as a respectable measure of reliability of WTP estimates. There are ways of improving the reliability of WTP studies and values. These include pre-testing of the survey instrument at development stage, adequate training of interviewers and statistical calibrations of WTP data such collected (Foreit & Foreit 2003; Mitchell & Carson 1989; Bowling 2001; Streiner & Norman 2008; Akter et al. 2007).

Scope (in)sensitivity in WTP is a measure of whether stated values discriminate between different ranges, sizes or proportions of valuation goods. Specifically, it is expected that WTP values would increase with higher quantities or scale of provision of the valuation good, until the point at which demand is satiated, *ceteris paribus*. However, while concerns have been raised about the possible scope (in)sensitivity of WTP values, like reliability, empirical evidence on this across sectors is mixed (Soto Montes De Oca & Bateman 2006). The assessment of scope (in)sensitivity has also been referred to in the CV literature as a test of construct validity, discussed in section 3.3.1. Challenges with reliability and scope (in)sensitivity of WTP values are acknowledged. However, these are beyond the scope of this thesis which focusses on the criterion validity of CV-WTP methods. The assessment of the criterion validity of CV-WTP is therefore presented in the next section.

3.4 The assessment of criterion validity in CV-WTP studies

Criterion validity has the greatest potential for offering a definitive test of a measure's WTP validity (Mitchel & Carson 1989). Assessment of criterion validity requires a criterion which is unequivocally closer to the theoretical construct than the measure whose validity is being assessed (Kane 2001). This criterion is also referred to as the "gold standard". This is similar to the equation of randomised control studies (RCTs) as the "gold standard" in clinical studies. However, in CV studies, there is no "gold standard" against which hypothetical values can be compared for the assessment of criterion validity. In the absence of this "gold standard", revealed preference studies have been conducted to generate values which are compared against the stated hypothetical values (Michell & Carson 1989, Murphy et al. 2005). The observed values from the revealed preferences are used as the criterion against which stated preferences are validated (Freeman, 1993). Revealed preference studies are used to derive "actual values". In this way, criterion validity is assessed. The different types of revealed preference studies against which to assess hypothetical WTP values are outlined in figure 1.1 and the methods are further described in Appendix 1.

As discussed in chapter 1, the total economic value comprises of both use and non-use values. Further, the TEV is assumed to incorporate the entire societal cost of any intervention. Of the five revealed preference techniques presented in figure 1.1, only the market prices method can be used to estimate directly both use and non-use values of a given good and are the closest comparator for criterion validity assessments. Actual market prices present the closest theoretical construct against which hypothetical values can be assessed, making this a criterion of central importance to CV studies (Mitchell & Carson 1989). The comparison of WTP values with actual market prices therefore would be the ideal criterion validity assessment. However, as discussed further in section 3.1, the CV method is often used in contexts where the market has either failed or is difficult to construct, as is the case with most public goods. Even where market prices are not available, simulated markets can be constructed and the amounts respondents in these markets pay or are willing to pay compared to the hypothetical CV markets for criterion validity assessment (Freeman III 2003; Mitchell & Carson 1989). Criterion validity assessments consider the extent of disparity between the hypothetical WTP and actual values. This disparity is also known as hypothetical bias.

The examination of hypothetical bias is an ongoing process in efforts to improve the CV method. Unfortunately, there is no widely accepted general theory explaining hypothetical bias that could provide guidance for an appropriate model specification (Murphy et al. 2005; Ajzen 2004). Researchers have tested different hypothesis to attempt to explain hypothetical bias. As discussed earlier, Streiner (2008) argues that hypothesis testing in validity assessments can be a continual, never-ending task. For a method which offers great potential for welfare maximization as the CV method does, this process is critical. In addition to explaining hypothetical bias, criterion validity assessments have been conducted to test different calibration methods of dealing with hypothetical bias. Some of the suggested calibration methods include:

(1) Ex-ante instrument calibration methods, aimed at reducing hypothetical bias at the survey design stage and; (List & Gallet 2001; Murphy et al. 2005).

(2) Ex-post statistical calibration methods which are aimed at correcting or reducing the stated WTP for hypothetical bias (List & Gallet 2001; Murphy et al. 2005).

While hypothetical bias has been thought to be influenced by the experimental protocols used in designing studies, there are still no generally accepted guidelines on the experimental protocols. Cummings et al., (1986) and the National Oceanic and Atmospheric Administration (NOAA) panel (1993) provided the earliest guidelines on the conduct of CV studies aimed at minimizing hypothetical bias. Cummings et al., (1986) suggested that: (a) respondents must be familiar with the good being valued; (b) respondents must have had experience with the good before the valuation exercise; (c) uncertainty in the hypothetical scenario description must be minimised – the outcomes and provision mechanisms and rules must be clarified to the respondent before the survey and; (d) the study must elicit WTP and not WTA.

The NOAA panel guidelines on the conduct of CV studies and the adjustment of values derived thereof to determine an accurate estimate of actual value included: (1) use of the dichotomous choice (DC) method in face to face interviews; (2) a minimum response rate of 70% and; (3) the conduct of sensitivity analysis showing how the size of the environmental program (and by extension any other benefit) would affect the mean WTP. In acknowledging the presence of hypothetical bias, the panel recommended that the mean hypothetical WTP should be reduced by 50% to adjust for this. Thus, a calibration factor (the ratio of hypothetical to actual mean WTP values) of 2 is used to correct the bias.

However, missing in the discussion on hypothetical bias and a likely contentious path is what levels of disparity between hypothetical and actual values are acceptable and, what is not acceptable. For instance, how close is 'close enough' for criterion validity to be demonstrated? Questions also abound about how close the hypothetical WTP and actual values should be elicited. Asked too close to each other might introduce recall bias while a long time-frame risks among other things, changes in the respondent's economic status, thus affecting their WTP value. Respondents might also totally forget about the hypothetical study.

3.5 Conclusion and chapter summary

In this chapter, the contingent valuation method has been elaborately discussed. The method has its origins in the environmental sector where the first practical valuations were conducted. Since then, the method has been used in several valuations across the sectors. The theoretical foundation of the CV method in welfare theory has also been discussed, and the design and conduct of such studies. With several applications of the method, guidelines have been conducted on the design and conduct of CV studies. The analysis of data obtained from CV studies has also been discussed.

Despite the apparent strengths of the CV method, the widespread application of the method in economic evaluation decisions has been limited. In addition to the questions related to the complexities of implementing such studies, the validity of estimates obtained using CV-WTP studies has been criticized. The use of the contingent valuation method in a large-scale assessment of the harm caused by the Exxon Valdez oil spillage in the USA highlighted the use of the method, while also generating heated debates. As a result, a panel comprised of prominent economists and led by Professor Kenneth Arrow (1921-2017), the first winner of the Nobel Prize in economics, was set up to appraise the validity estimates obtained using the method (Arrow et al, 1993). Following a phenomenal case, the panel concluded that the CV method "*produces estimates reliable enough to be a starting point of a judicial assessment of damage*" (ibid). The panel also offered a set of guidelines for the design and conduct of CV studies. The outcome of this panel rendered credibility to the method, increasing the use of the method.

The assessment of the different types of validity for CV-WTP studies has been conducted across sectors. In addressing the objectives of the thesis, the methods used in the assessment of validity were investigated. This was done through a systematic review of empirical studies. While an understanding of the theory is important, the application of the methods is more informative for decision making, including the development of methods such as CV-WTP. Further, to address the aims of the thesis, a critical assessment of what forms of validity have been assessed, and how this has been done will inform the empirical analysis in subsequent discussions. This is the focus of the next chapter.

Chapter 4 The Assessment of Validity in CV-WTP Studies: A Systematic Review

4.1 Introduction

In chapter three, the theoretical underpinnings of the CV-WTP method were discussed. Despite the apparent strengths of the method, concerns on the validity of the method limit its widespread application, especially in the health sector. The different types of validity were presented in the previous chapter. The aim of this chapter is to explore the assessment of the different types of validity and critically appraise the empirical evidence. A systematic review was conducted for this purpose and this is discussed. In contributing to the aim of the thesis of examining the criterion validity of contingent valuation WTP studies, the review highlights the challenges with the assessment of all types of CV-WTP validity. In the next section (4.2), the methods used in this review are presented. This is followed by a presentation and discussion of the results in section 4.3 and 4.4 respectively. In the conclusion in section 4.5, a case is made for further work on the criterion validity assessment in WTP studies.

4.2 Systematic review methods

An initial scoping search was carried out in June 2015 to establish the scope of literature on the validity of WTP methods and identify a possible search strategy. Four systematic reviews on CV methods were identified (Diener et al. 1998; Klose 1999; Yeung & Smith 2005; Yasunaga et al. 2006). None of these reviews explored the validity of CV methods. However, the search terms identified in these reviews were used to inform the search strategy used in this review. The term “validity[12]” and related terms were included to refine the search. The Prisma guidelines for the reporting of systematic reviews were followed for the narrative review (Moher et al 2009).

[12] A broad range of validity related search terms were used in recognition of the variety relating to this subject.

4.2.1 Databases

In September 2015[13] eight electronic databases (EconLit, TRID, MEDLINE, Embase, Web of Science, Psych info, CRD and CINAHL Plus) were searched from their inception to August 2015. The choice of the databases was guided by the scoping review discussed in section 4.2 above. Reference lists for relevant studies were also screened for additional articles.

4.2.2 Search strategy

In the absence of thesaurus terms on the subject, the following free text terms were used: Willingness to pay (Will*(or $) AND Pay or WTP), Willingness to accept (Will*(or $) AND Accept or WTA), Contingent valuation (Contingent Val* (or $) or CV), Hypothetical Value(s) (Hypothetical Value* (or $) Hypothetical Market*(or $)), Stated Preference(s) or Value(s) (Stated Preference*(or $) or Stated Value*(or $)). All the above terms were crossed with the term validity (Valid* (or $) or Construct Val* or Criterion Val* or Content Val* or Face Val* or Discriminant Val* or Convergent Val* or Theoretical Val* or Sensitivity AND Scope or Psychometr* or Psychological test*. Search strategies adopted suitably for each database (see Appendix 4 for sample Medline data base search strategy used). The results were managed using RefWorks reference management software.

4.2.3 Study inclusion criteria

Studies were included in the review if they:

1. Were conducted and reported in English[14] language;
2. Reported empirical WTP or willingness to accept (WTA) values and;
3. Assessed the validity (construct or criterion) of WTP or WTA.

4.2.4 Study selection

A three-stage process was used in the study selection process.

1. Titles and abstracts of the identified studies were reviewed for appropriateness and relevance according to the pre-determined inclusion criteria. Studies which did not meet the inclusion criteria were rejected at this stage. Where a definite decision could not be made from the title and/or abstract, the full paper was

[13] Searches were updated using automated monthly alerts. No new evidence has been generated on the methods used to assess validity since this time.
[14] The limitation to studies conducted in English was for feasibility (financial and time) purposes.

downloaded and reviewed to facilitate this decision. Where abstracts were not available, the full text of the paper was assessed.

2. Full text papers of potentially relevant studies were screened using the above criteria.
3. Finally, selected studies were categorized into the different types of validity tested. For some of the identified papers, authors' definitions of validity do not match the reviewer's definitions. These papers were classified based on the reviewers' classification of validity outlined in Appendix 3. Where a study assessed more than one type of validity the results for the different tests are presented separately.

4.2.5 Data extraction

A data extraction form was developed and used in this review (see Appendix 5). The form included questions on the broad context of the studies (geographical location), study type, sample size, aim and design, the study intervention (health or non-health) and intervention characteristics (goods or services and disease where applicable), respondent characteristics (sex, experience with the intervention) welfare measurement (measure, description of the intervention/ scenario descriptions, study perspective, WTP elicitation format, payment vehicle and time frame), validity assessment methods (such as hypothesis tested) and analytic methods (such as regression models and tests conducted), results including summary WTP estimates, respondent characteristics such as mean age and the authors conclusions on validity. A quality criterion on which to judge the selected studies was not available in the literature and therefore this was not done.

4.2.6 Data analysis

Given the heterogeneity in the studies identified from this review, only a narrative synthesis was conducted. Further, results are clustered by the validity type.

The review results are presented in the next section.

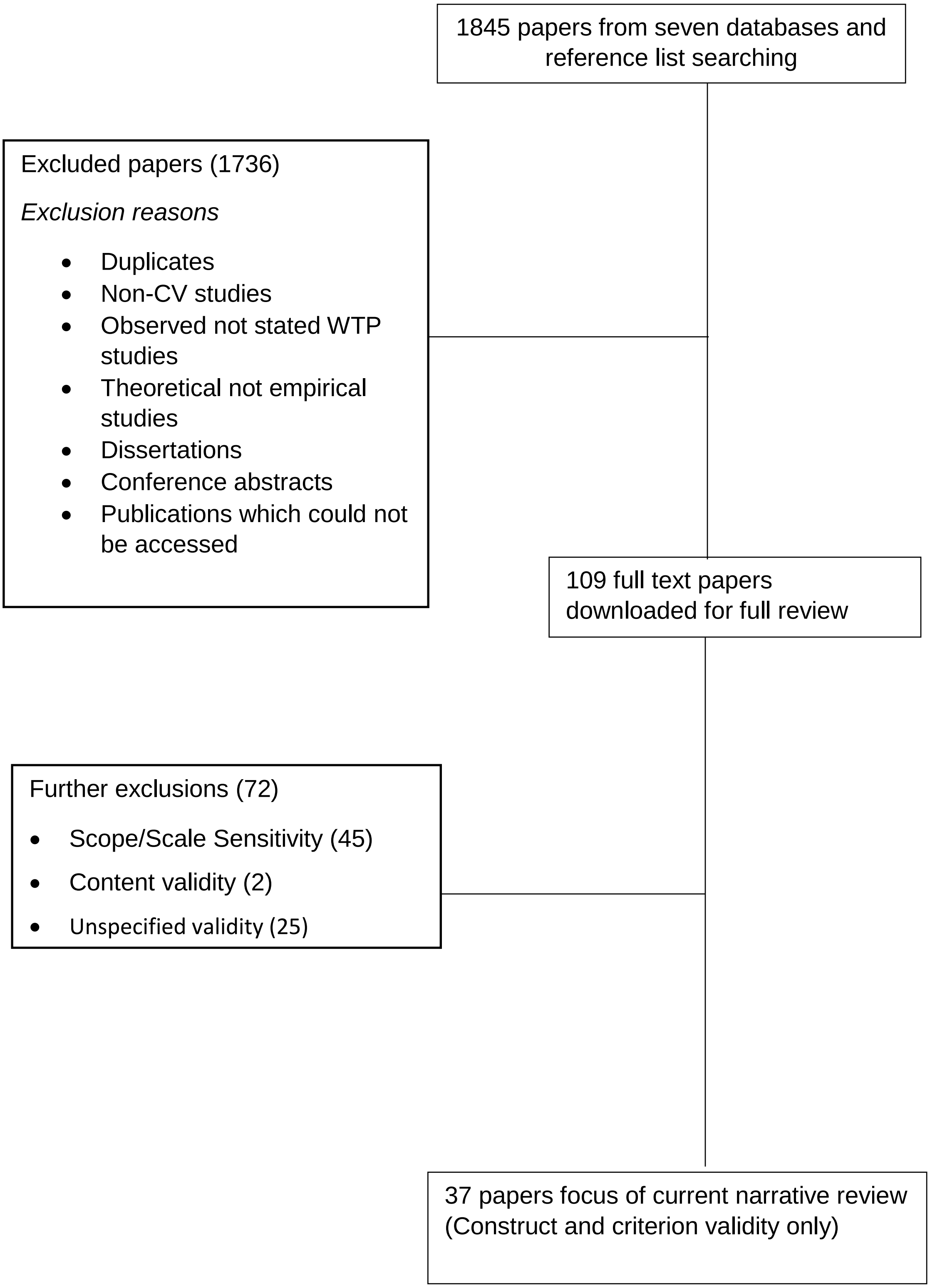
4.3 Review results

This section begins with a broad overview of the search results. Using a PRISMA diagram the flow of articles during the search process is presented. The findings on the methods used to assess the different types of validity are presented. A discussion synthesising the findings across the different types of validity concludes the section.

4.3.1 Overview of the database search

A total of 1845 abstracts were generated from the database search and reference list searches. Of these, 109 studies were downloaded and reviewed to facilitate the inclusion decision. Seventy-two of these were rejected based on the inclusion criteria with 37 articles left in the final selection. This flow of articles is illustrated in figure 4.1 while a list of all the included studies is provided in Appendix 6.

Figure 4-1: Screening and selection of review papers for narrative synthesis


1845 papers from seven databases and reference list searching
Excluded papers (1736)
Exclusion reasons
Duplicates
Non-CV studies
Observed not stated WTP studies
Theoretical not empirical studies
Dissertations
Conference abstracts
Publications which could not be accessed
109 full text papers downloaded for full review
Further exclusions (72)
Scope/Scale Sensitivity (45)
Content validity (2)
Unspecified validity (25)
37 papers focus of current narrative review (Construct and criterion validity only)

4.3.2 Background characteristics of the selected papers

Table 4.1 summarizes some key characteristics of the reviewed studies. Most of the studies were conducted in the USA (17) with 3 in the Netherlands, 4 in Australia; and one study each in Germany, United Kingdom, France, Mexico, Nigeria, Norway, Switzerland, Palestine, Sweden, Thailand, Canada and India. A multi-country study was conducted in Pakistan, Mali, Guatemala, and Ecuador.

Table 4-1: Characteristics of the reviewed studies

Attribute	**No. of papers***
Validity type	
Construct Validity	30
Criterion Validity	10
Sector	
Environment	21
Health	16
Other	3
Welfare measure[15]	
WTP	39
WTA	1
Study administration	
Face to face interviews	16
Mail surveys	11
Telephone interviews	7
Combined methods	6
Payment vehicle	
Out of pocket payments	23
Voluntary payments / donations	3
Tax / additional tax payments	7
Not stated	7
Payment frequency	
One-off payments	10
Annual payments	16
Daily payment	1
Not Indicated	13

*Some studies assessed multiple attributes hence the difference in figures.

[15] Welfare measures: These were defined as footnotes in section 3.1.3.

4.3.3 Findings on validity assessment

The review results are presented by the validity type as detailed in Appendix 3. The review summary focussed only on construct and criterion validity assessments as these involved analytical assessments relevant for the thesis topic. For each validity type that is discussed, an operational definition is provided, the methods used in the assessment of validity, including analytical tests and the general conclusions are provided.

4.3.3.1 Construct Validity

Construct validity examines whether hypothetical WTP estimates correspond to expected theoretical concepts (Klose 1999). This type of validity is examined using either convergent or divergent (discriminant) validity tests, both of which are defined and discussed in the following sections. In this review, construct and theoretical validity are taken to have the same meaning. Table 4.2 presents the key characteristics of studies examining construct validity.

Convergent Validity

Twenty-eight studies examined the convergent validity of CV-WTP estimates. Sample sizes in these studies ranged from 50 to 3,143 respondents. General and WTP question response rates were not indicated for eight and twenty-one studies respectively. In more than half the studies (17) of the studies, the authors defined the hypotheses that were tested. For many of the studies, this was defined as the presence or absence of differences in the estimates derived from the two samples used to test for convergent validity. Study hypothesis were either not defined or unclear for twelve of the studies. Approximately 40% of the reviewed studies did not report a clear conclusion on convergent validity. In those cases, the conclusion on convergent validity was deduced from the study.

Table 4-2: Key characteristics of studies examining construct validity

Author	Construct validity type (Convergent/ Divergent)	Sample Size	WTP/ WTA Elicitation format	Test	Author conclusion on criterion validity confirmed (√) not confirmed (x) not reported/ unclear
McCollum & Boyle, 2005	Convergent validity	1st stage experienced sample 900 (Dichotomous Choice (DC)*700; Open Ended (OE) 200); Non-experienced sample 600 (DC 500; OE 100) 2nd stage Residents (experience 900; Non-experienced 900); Non-residents (Experienced 100; Non-Experienced 100)	Dichotomous choice and Open-ended methods	Logistic regression	√: dichotomous choice format X: open ended question format
Rolfe & Dyack, 2010	Convergent Validity	890	Dichotomous choice for CVM	Regression techniques for CVM data and truncated negative binomial models for TCM data	x
Marjon et.al., 2008	Convergent Validity	CVM: 292 DCE: 100;	Discrete Choice Experiments; Open ended formats	Regression analysis	√
Boyle& Özdemir, 2009	Convergent Validity	2000		Test for the differences in coefficient and scale parameter estimates between different treatments	Mixed findings; convergent validity confirmed for one of the three treatments (hypothesis)
Vossler et.al., 2003	Convergent Validity	1209	Dichotomous choice with certainty and multiple bounded discrete choice	Parametric and nonparametric tests – comparison of the models	√
Champ et.al., 2006	Convergent Validity	First Wave. Videotape survey: 223, Phone 794; Second Wave. Videotape Survey:111, Phone 257	Open ended with follow up	Logit regression model	√

Author	Construct validity type (Convergent/ Divergent)	Sample Size	WTP/ WTA Elicitation format	Test	Author conclusion on criterion validity confirmed (√) not confirmed (x) not reported/ unclear
Sangkapitux et.al., 2010	Convergent Validity	First round: Face to face interviews: 562; Mail questionnaires: 860 Second survey round: Face to face interviews: 682; Mail questionnaires: 1150.	Dichotomous choice, Payment Card	Response probabilities modelled in a linear probit specification	√
Clarke, 2002	Convergent Validity	595	Bidding	Random utility model	x
Chambers et al., 1998	Convergent Validity	305	Payment Card	Tobit and Cragg regression techniques	√
Whitehead et. Al., 1998	Convergent Validity	1021	Dichotomous choice Polychotomous choice	Binary logistic regression	√
Severens et. al., 2000	Convergent Validity	84	Bidding	Spearman's Rank Correlation Coefficient and Stepwise logistic regression	x
Philips et al., 2006	Convergent Validity	1524	Payment Scale, Open ended method	Regression techniques	√
Barron et.al., 2004	Convergent Validity	62	Dichotomous Choice Bidding	Correlation and Regression	√
Onwujekwe & Uzochukwu., 2004	Convergent Validity	425	Open ended Binary with follow up	Heckman Selection models; Log OLS	
Bobinac et.al., 2012	Convergent Validity	1091	Payment Scale Bounded open ended	Logistic regression (including multivariate regressions)	x

Author	Construct validity type (Convergent/ Divergent)	Sample Size	WTP/ WTA Elicitation format	Test	Author conclusion on criterion validity confirmed (√) not confirmed (x) not reported/ unclear
Hoevenagel, 1996	Divergent Validity	1123	Payment Card	Parametric (t-test) and non-parametric (Mann-Whitney tests)	√
Smith, 2001	Discriminant validity	50	Open Ended	Non-parametric Wilcoxon paired comparison test	√
Onwujekwe et al., 2008	Convergent Validity	Bidding game: 261 Binary with follow up: 267 Structured Haggling: 273	Bidding game Binary with follow up Structured Haggling	Bivariate OLS OLS Multiple regression analysis	Overall results indicate some degree of validity. SH demonstrates greatest validity with BWFU demonstrating the least
Herriges et al.,1999	Convergent Validity	3143	Dichotomous Choice	Standard probit models	√
Kartman, Stålhammar & Johannesson., 1996	Convergent Validity	461	Dichotomous choice with open ended	Logistic regression	√
Lienhoop &Ansmann., 2011	Convergent Validity	591	Payment card	Logistic regression	√
Vossler &Watson, 2013	Convergent Validity	2000	Dichotomous choice	Logistic regression	√
Lew & Wallmo., 2011	Convergent validity	Total: 1,120 surveys 3 Species sample: 745 2 Species sample: 375	Stated preference choice experiments	Random utility maximization-based discrete choice econometric models	√
Montes de Oca & Bateman., 2006	Convergent Validity	1424	Dichotomous Choice	Probit Regression models	√

Author	Construct validity type (Convergent/ Divergent)	Sample Size	WTP/ WTA Elicitation format	Test	Author conclusion on criterion validity confirmed (√) not confirmed (x) not reported/ unclear
Blomquist & Whitehead., 1998	Convergent Validity	730	Dichotomous Choice	Logistic regressions	√
Camacho-Cuena et.al., 2004	Convergent Validity	Intervention: 76 Control: 124	Payment card Dichotomous choice	Regression analysis	√
Mataria et.al., 2004	Convergent validity	785	Payment Card	Tobit regression	√
Telser, Becker & Zweifel., 2009			Discrete choice	Regression analysis	
Veisten et.al., 2004		1019	Open ended Payment Card	Tobit regression	x
Foreit & Foreit., 2003	Convergent Validity	6683	Dichotomous choice Open ended Bidding game	Direct estimation techniques	√

Where both construct and convergent validity are reported in a paper, this is reported separately in the tables

Convergent validity is established if two or more different measurement techniques provide statistically indistinguishable estimates of the same theoretical concept (Carmines & Zeller 1979). In particular, convergent validity relates to whether the measure under consideration relates to other measures or constructs in a way that is predicted by theory (Clarke 2002). In the reviewed studies, convergent validity of WTP methods is assessed in four main ways:

1. Comparing WTP values obtained using different elicitation techniques;
2. Comparing (1) stated CV WTP estimates with values obtained using revealed preference techniques or; (2) estimates obtained using two stated preference techniques;
3. Checking for the agreement of WTP values with other known non-theoretical attributes such as experience and;
4. Checking the conformity of WTP values with expected theoretical constructs or economic theories which form underlying arguments for the validity assessments;
5. Examining other study design attributes such as WTP elicitation techniques.

The findings of the review are discussed by the above five broad assessment methods.

1. *Comparing WTP values obtained using different elicitation techniques*

Convergent validity was assessed by comparing WTP estimates obtained using different question formats such as dichotomous choice and open ended; multiple bounded discrete choice (MBDC) and dichotomous choice with follow up (Mccollum & Boyle 2005; Vossler, Ethier, et al. 2003; Onwujekwe & Uzochukwu 2004); dichotomous choice and polychotomous choice (Whitehead et al. 1998); and, open ended and binary with follow up formats (Onwujekwe et al. 2001). Where dichotomous choice questions were asked, convergent validity was checked by testing the equality of distributions using vectors of estimated parameters while the equality of mean values was tested for responses obtained using the open-ended format.

One study assessed convergent validity using three question formats: bidding game (BG), binary with follow up (BWFU), and the structured haggling (SH) technique in an assessment of WTP for mosquito bed nets (Onwujekwe et al. 2007). Based on the results from the study, the authors concluded that CVM was reasonably valid in the study settings and with different degrees of validity for the three question formats. The SH technique demonstrated greater validity with the BWFU least convincing when the three formats were compared (Onwujekwe et al. 2007).

Vossler (2003) compared corrected mean WTP estimates using two question formats, multiple bounded dichotomous choice (MBDC) and dichotomous choice with certainty. The authors concluded that mean WTP values obtained using the two formats corresponded, therefore confirming convergent validity (Vossler et al. 2003).

2. ***Comparing (1) stated CV WTP estimates with values obtained using revealed preference techniques or; (2) estimates obtained using two stated preference techniques***

Convergent validity has also been assessed by comparing estimates obtained from stated CVM WTP studies with those obtained through revealed preference techniques such as the travel cost method (TCM) (Clarke 2002; Rolfe & Dyack 2010). In his study, Clarke (2000) examined convergent validity by comparing WTP values for improved access to mammographic screening obtained using the contingent valuation (CV) and travel cost (TCM) methods (Clarke 2002). WTP estimates were obtained in a study estimating the recreational values associated with the Coorong on the Murray River in south-eastern Australia. Where WTP values obtained using two different preference elicitation methods such as the CV and TCM are compared to assess convergent validity, regression techniques are used to determine the variables that influence the decisions made by respondents (Clarke 2002). The coefficients of the variables are further examined and where concurrence is determined then convergent validity is confirmed, with the reverse holding true.

In their comparison, the confidence intervals (CI) for the mean WTP estimated using the CV methods were significantly higher than those of the TC models but these did not overlap. The CVM technique generated significantly lower estimates than the TCM and hence convergent validity was not confirmed in the study (Rolfe & Dyack 2010). However, convergent validity is confirmed in a different study using the same methods (TCM and CVM) (Lienhoop & Ansmann 2011).

Willingness to pay estimates were obtained using two stated preference techniques, CVM and discrete choice experiments (DCE) in one study (Marjon et al. 2008). Convergent validity was examined by comparing (i) the preferred attribute levels implied by each method and (ii) estimates of willingness to pay derived from each approach. The authors concluded that there was consistency in the estimates obtained using the two methods, confirming convergent validity.

3. *Checking for the agreement of WTP values with other known non-economic theoretical constructs*

Convergent validity assessments conducted in this way tested several hypotheses that do not have a foundation in economic theory. These included the exploration of attributes such as previous experience with a disease, condition or attribute of an intervention (Mccollum & Boyle 2005; Onwujekwe et al. 2007; Mataria et al. 2004), the subjective importance of an intervention to individuals (Severens et al. 2000), objective information about risk and behaviour (Philips et al. 2006), individual motivation towards an intervention (Foreit & Foreit 2003), health related quality of life measures of disease severity, pain and discomfort (Barron et al. 2004; Bobinac et al. 2012; Kartman et al. 1996) consequentiality (Vossler & Watson 2013), magnitude or scope of benefit or intervention (Lew & Wallmo 2011; Telser et al. 2008; Veisten et al. 2004; Camacho-Cuena et al. 2004) and resource quality information (Blomquist & Whitehead 1998).

In their study, McCollum and Boyle (2005) investigated the effect of respondent experience or knowledge in the elicitation of CV values as a test of convergent validity (Mccollum & Boyle 2005). The study recruited both respondents who had and those who did not have any prior experience with the valuation good. The authors sought to investigate whether valid responses would be obtained from those lacking experience. In the study, the underlying construct, ‘experience with the good subjected to valuation (direct hunting)’ was not needed for convergent validity to hold for dichotomous choice questions. However, this was needed for the open ended questions. Convergent validity was therefore confirmed for the dichotomous choice format but not for the open ended format (Mccollum & Boyle 2005).

In assessing the validity of WTP estimates against perceptions of risk and health benefit, authors Philips, Whynes and Avis (2006) concluded that WTP values behaved in a fashion consistent with that which would be expected in the market, supporting the validity of the method as used in the study context (Philips et al. 2006). Using similar methods in two different studies, authors established convergent validity of the WTP estimates when correlated with health related quality of life measurements of pain and discomfort (Barron et al. 2004; Kartman et al. 1996). However, in a different study convergent validity was not established when WTP estimates were correlated with health related quality adjusted life year (QALY) estimates (Bobinac et al. 2012).

4. ***Checking the conformity of WTP values with expected theoretical constructs or economic theories which form underlying arguments for the validity assessments***

In majority of the studies (n=23), the expected theoretical constructs or economic theories regarding the attributes compared with the hypothetical WTP or WTA values are defined. These form the underlying arguments for the validity assessments. The most common theoretical construct investigated in convergent validity assessments was income. Convergent validity was confirmed where WTP results conformed to theoretical expectations. For instance, at higher incomes, stated WTP was higher, *ceteris paribus* and vice versa. Various socio-demographic attributes, the majority of which were treated as proxies for income (e.g. education and employment status) were also assessed for expected convergence with hypothetical WTP/ WTA values (Chambers et al. 1998; Barron et al. 2004; Onwujekwe et al. 2007; Kartman et al. 1996; Soto Montes De Oca & Bateman 2006; Foreit & Foreit 2003).

In studies correlating WTP estimates with pre-determined socio-economic variables such as income, convergent validity was confirmed in eight studies (Philips et al. 2006; Barron et al. 2004; Chambers et al. 1998; Hergies et al., 1999; Soto & Bateman 2006; Blomquist & Whitehead 1998; Mataria et al. 2004; Veisten et al. 2004; Foreit &Foreit 2003) but refuted in one study (Onwujekwe & Uzochukwu 2004). Convergent validity was also not ascertained in a study correlating consequentiality with WTP estimates (Vossler & Watson 2013).

5. *Testing convergent validity by examining other study design attributes*

In two of the studies, convergent validity is ascertained by comparing WTP values obtained using different survey administration methods such as video tape and phone interviews (Loomis et al. 2006; Clarke 2002), and mail surveys and face to face interviews with citizen expert groups (Ahlheim et al. 2010). Multiple samples were used in the first two studies (Loomis et al. 2006; Clarke 2002). Mixed conclusions on convergent validity were deduced from the studies.

In one study, the authors investigated the framings of attribute based choice questions designed to measure the same theoretical value to assess construct validity (Boyle & Özdemir 2009). Still, in another study, authors assessed convergent validity by examining three study design issues: (i) the placement of the monetary stimulus (policy cost to respondents) in the sequence of attributes[16], (ii) the number of policy alternatives respondents are asked to consider in choice questions[17] (two versus three) and, (iii) the inclusion versus exclusion of a status-quo alternative in choice questions[18] (Boyle & Özdemir 2009). The authors conclude that the placement of the monetary stimulus and the inclusion/exclusion of a status-quo alternative in the choice questions did not affect the respondents' estimates of preference parameters but there were significant differences between questions with three versus two alternatives.

Divergent or discriminant validity

This type of validity was examined in two studies (Hoevenagel 1996; Smith 2001). The studies are briefly described and the findings on divergent validity discussed.

In their study, Hooevenagel (1996) investigated the extent to which the WTP method produces different values in situations for which economic theory claims differences (Hoevenagel 1996). The authors used perfect and regular embedding in WTP estimates as the constructs underpinning the divergent validity tests. The WTP

[16] In investigations of hypothetical valuations involving cash transactions, hypothetical bias has been observed. However, how the placement of the monetary question influences hypothetical bias is not yet known (Boyle & Ozdemir 2009).

[17] Evidence from empirical assessments suggests that complex choices are associated with higher hypothetical bias (Boyle & Ozdemir 2009).

[18] The inclusion of a status-quo alternative maximizes the options a respondent would have to an evaluation question (e.g. Yes, No, Don't know or Unsure). Failure to include this leads to incorrect estimations of the welfare evaluation (Boyle & Ozdemir 2009).

question was elicited in this study using the payment card method. Study hypothesis are stated and both parametric (t-tests) and non-parametric (Mann Whitney tests) tests used to investigate divergent validity. Given the study results, the authors concluded that the WTP method could discriminate between different environmental goods in a way that corresponds to the pre-determined economic axioms, thus confirming divergent validity (Hoevenagel 1996).

In the second study, discriminant validity was examined by assessing the relative sensitivity of WTP and time-trade-off (TTO) techniques to changes in health status (Smith 2001). Utility weights were determined using the TTO technique while WTP estimates were obtained using an open-ended question format. The constructs to be assessed were determined apriori and the Wilcoxon paired comparison non-parametric tests used to assess the difference between the two samples. The authors concluded that the WTP method was more sensitive than TTO in distinguishing between dimensions of health at the same nominal level of health status. The WTP was also more sensitive to differences in quality of life between different levels of health within each dimension, confirming divergent validity of the method (Smith 2001).

4.3.3.2 Criterion validity

Criterion validity tests are also referred to as tests of external validity of WTP in the literature (Hergies et al. 1999). Criterion validity is assessed through predictive or concurrent validity assessments. A total of 10 papers identified in the systematic review assessed criterion validity. The papers specifically assessed concurrent validity of WTP and are referred to in the discussion as criterion validity assessments. Key characteristics of these studies are summarised in table 4.3.

Sample sizes in these studies range from 60 to 430 respondents. In four of the studies, the authors set out hypotheses which were tested using the different methods (Loomis et al. 2009; Loomis et al. 1997; Vernazza et al. 2015; Muller & Ruffieux 2011). Study hypotheses were not clear in six of the studies (Loomis et al. 1996; Vossler & Kerkvliet 2003; Camacho-Cuena et al. 2004; Johnston 2006; Bhatia & Fox-Rushby 2003; Vossler, Kerkvliet, et al. 2003). The dichotomous choice WTP elicitation format was used singularly or in combination with other methods in seven of the studies (Vossler & Kerkvliet 2003; Camacho-Cuena et al. 2004; Loomis et al.

2009; Loomis et al. 1997; Johnston 2006; Bhatia & Fox-Rushby 2003; Vossler, Kerkvliet, et al. 2003). Three studies used the bidding format (Bhatia & Fox-Rushby 2003; Vernazza et al. 2015; Muller & Ruffieux 2011) while the open ended format was used singularly or in combination with other formats in three studies (Loomis et al. 1996; Loomis et al. 1997; M R Bhatia & Fox-Rushby 2003).

Criterion validity was primarily assessed through comparisons of WTP values obtained from hypothetical surveys of WTP and actual values obtained through laboratory experiments, also referred to as simulated market experiments (SMEs). Other assessment methods included tests of consequentiality and comparison of WTP estimates to; voting behaviour or purchase behaviour following price changes.

Criterion validity assessments were summarised as ratios between hypothetical WTP and actual values or summarised descriptively. In two of the studies assessing criterion validity, the authors did not present a conclusion (Loomis et al. 2009; Loomis et al. 1997) while this was unclear in two studies (Vossler & Kerkvliet 2003; Muller & Ruffieux 2011). These findings are summarised below.

1. *Comparisons of hypothetical WTP and actual values*

Hypothetical and actual WTP values were compared in six studies (Vossler & Kerkvliet 2003; Loomis et al. 2009; Loomis et al. 1997; Johnston 2006; M R Bhatia & Fox-Rushby 2003; Vernazza et al. 2015) to determine criterion validity. Hypothetical WTP values were elicited using surveys and laboratory experiments while observations of real cash market transactions in a laboratory setting (SME) were used to elicit actual WTP values in the studies.

In these studies, criterion validity was tested in the following ways which depended on the WTP elicitation format: (1) comparison of confidence intervals about estimates of hypothetical WTP and actual values; (2) determining whether the odds of agreeing to pay the bid amount are equal between the hypothetical and actual dichotomous choice treatments (3) regression analyses to determine whether the coefficients estimated are similar and;(4) calculating the mean difference between hypothetical WTP and actual values.

Table 4-3: Key characteristics of papers testing criterion (external) validity

Author	**Sample Size**	**WTP Elicitation format**	**Test**	**Author conclusion on criterion validity** **confirmed (√)** **not confirmed (x)** **not reported/ unclear**
Loomis et al., 2009	290	Dichotomous choice	Logistic regression models	Not reported
Loomis et.al., 1997	Open ended real cash WTP sample: 32 Open ended hypothetical WTP sample: 33 Dichotomous choice CVM: 56	Dichotomous choice Open ended	Comparisons of confidence intervals about estimates of hypothetical and actual WTP	Not reported
Johnston, 2006	430	Dichotomous Choice	Random utility model	x
Bhatia & Fox-Rushby, 2003	300	Bidding with open ended Dichotomous choice	Analysing the distribution of deviations between hypothetical and observed behaviour	x
Vossler &Kerkvliet., 2003		DCF; DKF; IRF	Logistic regression	Not clear
Muller&Ruffieux., 2011	128	Bidding in vickery auctions	Estimating mean differences in WTP values derived from the different auctions	Not clear
Vossler et.al., 2003	465	Dichotomous choice	Logistic regression	Not clear
Loomiset.al., 1996		Open ended	OLS regression	x
Camacho-Cuena et.al., 2004	Intervention: 76 Control: 124	Payment card Dichotomous choice	Regression analyses	v
Vernazza et.al., 2015	60	Bidding card	Ordinary least squares regression	x

The ratio of hypothetical WTP to actual WTP values using the different methods was considered high at factors ranging from 1 of 7.1 in two of the studies (Loomis et al. 2009; Loomis et al. 1997), and criterion validity was therefore not confirmed. In a study comparing hypothetical and referendum values the author reports criterion validity as limited[19] (Johnston 2006). In two studies comparing hypothetical WTP values with demand for commodities, authors conclude that the WTP method did not demonstrate criterion validity (Bhatia & Fox-Rushby 2003; Vernazza et al. 2015). In a study validating hypothetical WTP for a recyclable product, the authors conclude that the CVM method was externally valid in the study population (Camacho-Cuena et al. 2004).

2. *Other assessments of criterion validity*

Tests of consequentiality are used to assess criterion validity in one study (Loomis et al. 2009) while two studies evaluated criterion validity by comparing WTP values elicited using different WTP/WTA question formats such as dichotomous choice and open ended methods (Loomis et al. 1997; Vossler & Kerkvliet 2003). One study compared WTP values elicited using three WTP elicitation formats to actual voting behaviour in a referendum (Vossler & Kerkvliet 2003). In a different study, the impact of price signals as observed from WTP bids before and after the introduction of price tags through auctions was used to assess criterion validity (Muller & Ruffieux 2011). Criterion validity is reported as strong[19] in this study. The analytic tests conducted to assess criterion validity in these studies is unclear.

4.4 Discussion on the assessment of the validity of CV-WTP methods

The reviewed studies span across different sectors with majority of the evidence generated from the environmental sector. Given the small number of studies generated from the search, the number of studies identified from each sector is even fewer.

The validity terminology used in the studies is varied, with some studies using the terms construct and criterion validity interchangeably. Construct validity has been referred to in some papers as criterion, theoretical and convergent validity while criterion validity has been referred to as construct and external validity. Despite the different terms, the methods describing the process of testing for either construct or

[19] The meaning of “limited” and “strong” criterion validity was not clear from the studies.

criterion validity are clear, and this enabled a reclassification of the studies based on the reviewers' validity definitions.

Face and content validity assessments provide the crucial first test for any CV study. These two validity tests assess whether the CV study asked the right questions in a clear, understandable, sensible and appropriate manner based on which a valid WTP/WTA estimate would be obtained. However, only two of the papers identified in this review indicated having conducted content validity tests. Details of how the content validity tests were conducted are not provided, making it difficult to judge the instruments used to examine construct and criterion validity. Low content and face validity of CV estimates could lead to high rates of general survey and WTP question non-response, and therefore hypothetical bias.

General survey and WTP/WTA question non-response rates are not reported in majority of the studies. Similarly, plausible reasons for the non-response for the WTP question are generally not provided. General and WTP question non-responses reduce the sample size, which affects the analysis and the conclusion drawn from the studies. The effect of non-response values on the sample sizes is not reported in majority of the studies. The way missing data is handled in the analysis is also not reported in majority of the studies further limiting the interpretation of the results.

Socioeconomic and demographic characteristics of study respondents such as income would be regarded basic construct validity tests in a WTP study (Arrow et al. 1993). While respondent income is asked in majority of the studies, correlations of income with WTP are often not reported. In the absence of a "gold standard", assessment of criterion validity in majority of the studies involves the use of actual willingness to pay as the criterion against which hypothetical WTP values are assessed. In most of these studies, actual willingness to pay is elicited in laboratory conditions. There is limited evidence on the validity of estimates obtained from such experiments when used as the criterion against which hypothetical WTP values are elicited.

Respondents' choices in these scenarios often have no direct financial consequences for them. The effect of consequentiality on hypothetical and actual WTP values elicited using these experiments remains the subject of ongoing research (Vossler, 2003). In other studies, different WTP elicitation formats have been compared to ascertain criterion validity of WTP. As with the laboratory experiments, the evidence on the validity of the different question formats is limited too. The evidence points to the potentially different effects of elicitation formats on hypothetical bias. In particular, open ended methods have been associated with higher hypothetical bias compared with discrete choice methods (Onwujekwe et al., 2004).

Different methods and tests are used to ascertain construct and criterion validity. Procedural invariance is assumed in majority of the studies or where this has been assessed, it is not reported. There is no consensus across sectors on the most appropriate methods or tests to use in determining validity. Based on the reviewed studies, both construct and criterion validity do not seem to depend on study attributes such as sample size, setting, WTP elicitation format, survey administration or tests. Rather, study findings are specific to the study setting, sector and valuation good or service. Given the limited number of studies and the even fewer number of validity studies in the different sectors, the generalizability of the results cannot be ascertained. This is compounded by the very specific nature of CV studies and hypothetical scenarios.

In reviewing the evidence, it is possible that the search terms utilised did not capture all the relevant articles for this synthesis. The diversity in the terms used for validity further complicated this search. However, as the focus was on summarising the evidence on the methods used to assess validity and highlight the common types of validity assessed, the results can be regarded as sufficient for this purpose.

4.5 Conclusion and chapter summary

Overall, the evidence on the construct and criterion validity of WTP methods from the reviewed studies is mixed. In particular, no trends are noted on either the methods used in the assessments or conclusions on validity. Even with careful consideration in the design and implementation of studies, many details which could explain study results further and enhance understanding on conclusions are not reported. Further, the limited number of studies on the subject makes it difficult to draw any firm conclusions on the validity assessment methods. This is especially so for criterion validity assessments with only 10 studies identified.

As discussed in the introduction to this thesis, a long-standing criticism of CV methods is that stated willingness to pay estimates may be a poor indicator of actual WTP (Loomis et al. 1996; Carson et al. 1996). The correlation between hypothetical WTP and actual values relates to the assessment of criterion validity. As discussed in the previous chapter, the assessment of criterion validity would involve comparison of hypothetical values to a “gold” criterion. However, the lack of a “gold standard” against which to evaluate CV studies presents a challenge for further criterion validity assessments. The assessment of criterion validity has therefore been conducted in different ways which then complicates the interpretation of conclusions thereof. The lack of guidelines for the conduct of validity assessments in CV-WTP methods complicates this further.

This chapter has highlighted the variety in the methods used in the assessment of construct and criterion validity in CV-WTP studies. Notably, using clearly defined systematic review criteria, very few criterion validity assessments have been conducted across the sectors (10). The diversity in the methods used to assess criterion validity in these studies does not permit a quantification of the magnitude of hypothetical bias. It is also likely that some studies were not captured given the strict systematic review search strategy. In attempts to further establish the evidence on how criterion validity assessments have been conducted and the magnitude of hypothetical bias, earlier reviews were sought. In this rigorous process, no limits were applied to the manner in which criterion validity was assessed. In addition to establishing the evidence on criterion validity, the reviews informed the design of a more focussed systematic review of empirical studies assessing criterion validity. These reviews are presented in the next chapter.

Chapter 5 Criterion Validity Assessment in CV WTP Studies: Systematic Reviews of the Evidence

5.1 Introduction

As discussed in the previous chapters, the CV method remains a subject of ongoing criticism, with critics' primary concern being that the method lacks criterion or external validity. External validity is also referred to as criterion validity and is the primary focus of this thesis. The systematic review on the methods used to assess the validity of CV-WTP identified very few empirical studies assessing the criterion validity of CV-WTP. However, the review also illustrated the diversity in the terms used to define the types of validity, including criterion validity. Further, the variety in the assessment methods was also established. In this chapter, two reviews conducted separately are presented. The first is a summary of reviews conducted on criterion (external) validity. This review summarises previous synthesis of criterion validity assessments. Based on the findings, a justification is made for the systematic review of empirical studies assessing criterion validity, which is also presented in this chapter. The summary also helps in developing an exhaustive search strategy for the second systematic review. The chapter is structured as follows: in the introduction, a broad discussion on criterion validity assessment is provided. This is followed by a discussion of the methods of the review of reviews, the results of the search and a discussion. Finally, the systematic review of empirical studies assessing criterion validity is presented and discussed.

5.2 Overview on criterion validity assessment

In chapter 3, the economic theory underlying welfare measurement using stated preference techniques was presented. The potential of the method to capture the total economic value, compared to other economic evaluation techniques was discussed (Mitchell & Carson 1989; Freeman III 2003). However, there is no widely accepted theory of how people respond to questions about their WTP when it is hypothetical (Murphy et al. 2005). Much of the literature suggests that hypothetical values tend to overestimate actual values (evidence of hypothetical bias) and this has remained the subject of ongoing criticism of the method.

Primarily, criterion validity assessments aim to determine whether hypothetical bias exists; and if or where it does, the magnitude of this disparity. Assessments also seek to establish the likely causes of the hypothetical bias based on which the CV method can be improved. Criterion validity tests conducted so far have yielded mixed results. While most of the studies have concluded that the CV method is not criterion valid, other studies have demonstrated this validity in assessments of different types of goods. Following the seminal work of Bohm (1972) on the assessment of hypothetical bias in WTP studies, several researchers have conducted similar assessments to date with the last published record being in 2016 (Ryan et al., 2016).

Authors have summarised both qualitatively and quantitatively, available WTP criterion validity assessments conducted using the above methods over time. In the next section, a summary is provided of the reviews of criterion validity assessments published to date. The summary highlights methodological findings and issues for consideration in further criterion validity assessments.

5.3 Review of reviews on the external (criterion) validity of WTP

5.3.1 Review aims

The purpose of this review of the reviews was to:

1. Establish the current view on stated and revealed preference methods and agreement on the criterion validity of CV-WTP methods;
2. Identify and categorise the types of revealed preference data that have been used to evaluate criterion validity and the justifications given for the use of these methods;
3. Determine how criterion validity has been examined and the justification for the methods including:
    a. Types of estimates included in the analysis and the justification for the selected estimates;
    b. Analytic methods used, including both regressors and regressands. This includes establishing any implied theoretical position.
    c. How multiple estimates of WTP/WTA values have been handled in the analysis
    d. Justifications for excluding any estimates;

4. Identify potentially useful methods to inform a systematic review and meta-analysis of empirical studies assessing the criterion validity of WTP methods including:
    a. Study inclusion criteria used and their justification;
    b. Categorisation of included goods or services e.g. class of goods;
5. Identify potential gaps in the methods used in the analysis of stated and/ or revealed preference data. This will further show how a systematic review and meta-analysis of empirical studies presented later advances current knowledge on the subject.

5.3.2 Review methods

A focussed review of all reviews and meta-analyses on the criterion validity of WTP methods published to-date was conducted in December 2015[20]. A systematic search was conducted in eight electronic databases (EconLit, TRID, MEDLINE, Embase, Web of Science, Psychinfo, CRD and CINAHL Plus). There were no time restrictions employed during the search process. In addition, reference lists of key papers identified during an earlier systematic review of validity assessments (chapter 4) as well as initial papers identified during this review was conducted.

The search terms used with the database searches were similar to those used in the review of validity assessments (Appendix 7). Additional terms “Review”, “Meta-analysis”, “Synthesis” were included to focus the search on reviews.

Papers were included if they:

1) Were reviews of empirical studies;
2) Conducted in English and;
3) Involved a comparison of hypothetical and actual values without any restriction on the techniques used.

5.3.3 Review results

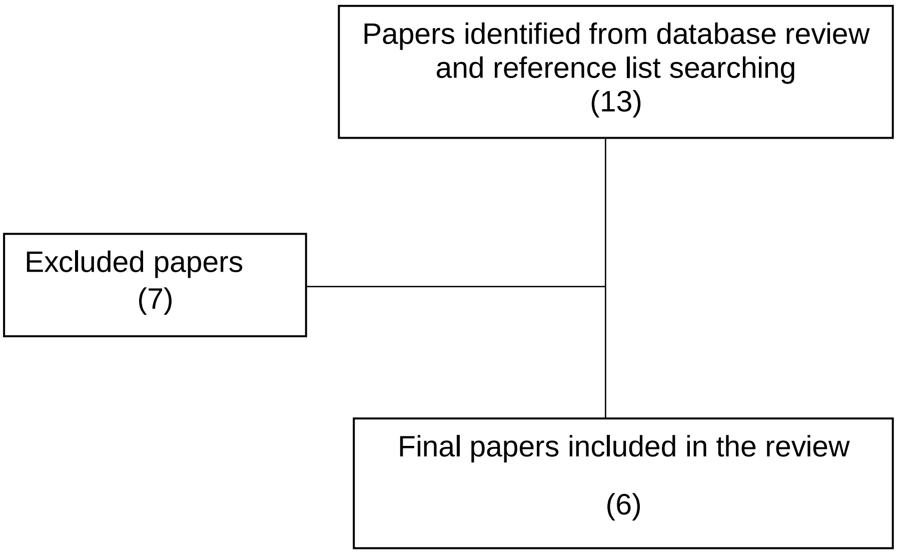
Figure 5.1 illustrates the flow of articles identified during this review.

A total of six publications were identified (Carson et al. 1996; Harrison & Rutström 2008; Liljas & Blumenschein 2000; List & Gallet 2001; Little & Berrens 2004; Murphy et al. 2005) (see Appendix 8 for list of included studies). Four of the reviews included

[20] An automated monthly search alert set up in December 2015 has not returned any new systematic review on the criterion validity of CV-WTP.

a meta-analysis. Little and Berrens (2004) expanded the work done by List and Gallet (2001). However, the meta-analysis by Murphy et al (2005) is a re-analysis of the earlier meta-analysis by List and Gallet (2001) and therefore in a strict sense, there have been only five (5) criterion validity assessment reviews (and three meta-analysis) published to date. A summary of the results is provided in the next section, in chronological order. Detailed results (also in chronological order) are provided in Appendix 9.

Figure 5-1: Flow of articles during literature search process


Papers identified from database review and reference list searching (13)
Excluded papers (7)
Final papers included in the review (6)

5.3.3.1 Findings from the reviews

In their review, Carson et al (1996) identified 83 studies with a total of 616 comparisons of hypothetical and actual values. The hypothetical values obtained using stated preference techniques were compared with actual vales obtained using any revealed preference technique. The values obtained were not meta-analysed owing to incomplete reporting of necessary details, but a meta-regression was conducted. The review authors established that CV/RP ratios of >2.0 comprise only 5% of the complete sample and only 3% of the weighted sample in their review. Based on this, they argue that the suggested calibration of WTP values as suggested by the NOAA panel (Arrow et al. 1993) of 2 is unwarranted.

The second review was conducted by Harrison and Rutstrom in 1999 but was not published until 2008. In this narrative review, the authors clustered 35 studies based on the type of good (public or private) and two broad elicitation techniques (open ended and dichotomous choice). Based on their findings, the authors confirm the presence of hypothetical bias. They also identified several design attributes such as the elicitation format, subject pools and the type of good (public or private) as potential drivers of hypothetical bias. The authors also highlighted the treatment of different response categories (e.g. “yes”, “definitely yes”) and the impact of these on the conclusions on hypothetical bias. The authors argued against claims of increased hypothetical bias with certain response formats and present a case for certain calibration methods to correct for this.

In the third review, Liljas, Bengt and Blumenschein (2000) summarised criterion validity assessments clustered around three elicitation methods: (1) Open ended WTP questions, (2) Auctions and, (3) Dichotomous choice questions. The authors also summarised the evidence on hypothetical bias for studies that had calibrated responses to correct discrepancies between hypothetical and actual WTP. Based on their analysis, the authors concluded that substantial hypothetical bias existed. They argue that this bias is good specific and not general. Further, the authors observed that the hypothetical bias might be related to poor study designs rather than a lack of validity of the CV method as had been previously argued. Like Harrison and Rutstrom (1999), the review authors argue that the certainty of a hypothetical “yes” response may be an important predictor of a real “yes” response in WTP studies. This therefore suggests the possibility of calibrating hypothetical WTP responses to better match real valuation.

In 2001, List and Gallet reviewed the evidence on a range of mixed goods to provide evidence pertaining to the effects of various experimental protocols on the observed calibration factors. The authors find that the average person exaggerated their hypothetical WTP across a broad spectrum of goods with vastly different experimental protocol, with only a few exceptions to this. Considering that the relationship between real and hypothetical values may be specific to experimental protocols, the authors further investigated, through regression models, the impact of a wide range of variables including: study setting, type of good, WTP/WTA, within or between subject comparisons and elicitation methods. The authors find that study subjects overstate their hypothetical WTP by a factor of at least 3. Despite this, the discrepancy between the minimum and maximum reported calibration factors is relatively small. The meta-regression shows that hypothetical bias is not affected by the study setting or whether within or between comparisons are made. However, regression analysis identifies the elicitation methods, WTP/WTA and the type of good as having a significant effect on hypothetical bias. The authors therefore conclude that certain experimental protocol influence the deviations in hypothetical and actual WTP values.

In the fifth review, Little and Berrens (2004) expand the original meta-analysis by List and Gallet (2001). More data is included in this review, with additional variables considered such as the inclusion of referendum formats, certainty corrections and cheap talk scripts. The review authors also explore the effect of weighting and clustering techniques since for some of the empirical studies, individuals produce multiple observations. The authors maintain the same analytical methods used by List and Gallet (2001). The review generated mixed results, when compared with the results obtained from the earlier review by List and Gallet (2001). With the expanded dataset, Little and Berrens confirm the findings in the List and Gallet (2001) analysis, that first price sealed bid auctions reduce hypothetical bias. However, contrary to the results of the initial analysis (List and Gallet, 2001), the review authors found that hypothetical bias is reduced with the use of public goods, referenda and with certainty corrections. Authors reaffirm the conclusions in earlier reviews about the effect of study designs on hypothetical bias.

In the last identified review and meta-analysis, Murphy et al (2005) re-analysed the meta-analysis conducted by List and Gallet (2001). They used refined criteria and found that the magnitude of hypothetical bias was statistically less for WTP compared to WTA, private compared to public goods and, the use of first price bidding methods as opposed to vickery second price auctions. Based on their analyses, Murphy et al (2005) concluded that hypothetical values are the best predictors for actual value and the use of calibration techniques could effectively reduce hypothetical bias. The review authors also observe that hypothetical bias is sensitive to model specification, a lack of variability in the data and the treatment of extreme values.

5.3.4 Discussion and conclusion on the reviews of criterion validity

The first review of criterion validity studies covered in this summary was published in 1996 with the last seven years later, in 2003. The reviews cover studies published from 1972 to 2003. While several criterion validity assessments have been conducted and published since 2003, there doesn’t seem to be record of any other review or meta-analysis of criterion validity studies for the last 13 years.

All the reviews defined hypothetical bias as the difference between hypothetical and actual values. Hypothetical values obtained using contingent valuation (CV) methods are compared with revealed preference (RP) techniques in all the reviews. Except for one (Carson et al., 1996), the reviews use actual values obtained using simulated markets, to compare against hypothetical CVM values. In their review, Carson et al., (1996) include studies using the travel cost, hedonic pricing, averting behaviour and actual values RP techniques. In all the reviews, there is no limitation on the type of good valued and while this is not explicitly stated in most, studies included are exhaustive in coverage up to the year of the publication (except where further criteria are applied to refine the included observations). Further, for all the reviews, multiple estimates of hypothetical and actual values from individual studies have been included. Four of the six reviews include a meta-analysis of the values obtained.

The effects of different experimental protocol 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 with the effect of these mixed across the reviews. For instance, Liljas, Bengt and Blumenschein Karen, (2000) concluded that hypothetical bias does not depend on the elicitation method while List and Gallet (2001) established an opposite effect. The findings on the effect of the type of good – whether public or private – on hypothetical bias is mixed with some reviews establishing a significant effect (List and Gallet, 2001) and others concluding that this has no effect (Murphy et al., 2003; Little Joseph and Berrens Robert, 2004). The experimental design (whether between-subject or within-subject) has been shown to have no effect on hypothetical bias (List and Gallet (2001). Finally, with only two studies conducted in the health sector in the included reviews, the evidence on criterion validity is primarily based on studies in the environment sector.

The results of these reviews have been mixed with the majority confirming the presence of hypothetical bias. Some of the review authors further observed that the large differences in real and hypothetical WTP might be related to poor study design rather than to lack of validity for the CV method. Most of the reviews also recognised that instrument and statistical calibration methods could be used to reduce hypothetical bias (Harrison & Rutström 2008; List & Gallet 2001; Little & Berrens 2004; Murphy et al. 2005).

Different regression models were used in the reviews to investigate the direction and significance of various independent variables on hypothetical bias. As a comprehensive theory of hypothetical bias has not been developed, the models are specified variedly, with latter studies borrowing ideas from former studies and including or adjusting some variables to improve the models. The fact that there exists a great variation in the studies included in the quantitative reviews also highlights the difficulties inherent in the analysis. This variation also makes it difficult to make conclusions on the criterion validity of CV-WTP based on the available results. While this applies across the sectors, the complexity is compounded by the few studies available in sectors such as health. This therefore limits the depth of analysis that can be conducted to isolate differences attributable for example, to sector and welfare measures. With no guidelines on the conduct and analysis of criterion validity assessments, and the wide scope of studies included in this review, such variety in methods and results is not surprising. However, further investigations of the effect of the different methodological approaches to assessing the criterion validity of CV-WTP might offer some insight on study attributes that might influence hypothetical bias.

Notably, the last synthesis of studies assessing the criterion validity of WTP was conducted more than one decade ago (2003). Since then, several criterion validity assessments have been conducted. These have tested different experimental protocol across sectors with varying results. With more studies comparing stated and actual values conducted, this allows for the effect of different methodological attributes on hypothetical bias to be evaluated. This therefore justifies the need for another systematic review updating the last one by Murphy et al. (2003). An updated review might also highlight improvements in both the conduct and analysis of criterion validity assessments and may derive important methodological findings regarding CV-WTP methods. Significantly, this review further demonstrated the diversity in the terms used to describe criterion validity assessments. This helped in designing a focussed review of empirical assessments of criterion validity.

A systematic review of empirical studies assessing the criterion validity of CV-WTP methods, with refined inclusion criteria is discussed in the next section. The methods used to conduct the systematic review are presented first and this is followed by the detailed results of the search. In the discussion, the key methodological issues in the design of hypothetical and actual surveys assessing the criterion validity of WTP are presented. The chapter concludes with a discussion which points to the need for guidelines in the reporting of CV studies and particularly criterion validity.

5.4 Systematic review of empirical studies assessing the criterion validity of CV-WTP methods

The review follows the PRISMA guidance on methods for conducting and reporting of systematic reviews (Moher et al. 2009). A protocol for this review was not registered.

5.4.1 Methods

5.4.1.1 Literature search strategy

Eight electronic databases (EconLit, TRID, MEDLINE, Embase, Web of Science, Psychinfo, CRD and CINAHL Plus) were searched from their inception to December 2015[21]. Valuation terms (willingness to pay, willingness to accept, contingent valuation, 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) to search each database similarly (see Appendix 10 for a sample search strategy). In addition, reference lists, citation and author searches were conducted to identify additional papers. The methods used in this review were informed by the previous reviews on criterion validity presented in section 5.3. Results were managed using Mendeley reference management software.

5.4.1.2. Study Selection criteria

All titles and abstracts, and full papers when in doubt, were double-reviewed and

[21] An automated monthly search alert was set up in December 2015. None of the publications identified through the alert fit the inclusion criteria for this systematic review.

studies included if they were:

1. Conducted and reported in English;
2. Assessed criterion validity of WTP/WTA;
3. Included direct WTP elicitation methods (CVM) 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.

5.4.1.3. Data Extraction

Data was extracted using a standard template in MS Excel (see Appendix 11). A second reviewer double extracted data for a randomly selected 10% sample. Disagreements were resolved through discussion, with any implications followed through to all other papers. Extracted data included background characteristics (e.g. country, validity terminology used, good valued), survey design (e.g. welfare perspective, elicitation format and pre-specified values for both hypothetical and actual WTP surveys, payment vehicle, mode of administration), 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). All these study attributes are discussed in Appendix 12. A quality rating was not employed as no agreed criteria exist on criterion validity assessments.

5.4.1.4. Statistical analysis

A narrative and quantitative summary of the methods used in the comparisons of hypothetical WTP and actual values from the reviewed studies and findings is provided. As the majority of the papers identified report multiple comparisons of the hypothetical and actual values, the results are reported by the comparisons rather than by paper or study. This is to allow for the use of all the estimates and hence a larger dataset for the analysis. For all comparisons, WTP estimates for hypothetical and actual data are matched as pairs, when provided, and compared as a ratio (for mean values) and as odds ratios (for percentage summaries). All analyses were conducted using Stata14.

The entire dataset of included comparisons is used in the narrative summary. The comparisons of hypothetical and actual values in terms of background characteristics, survey design, study methods and results are summarized using counts, descriptive statistics, 2 by 2 tables, and box and whisker plots. Further analysis and comparisons of hypothetical WTP and actual values are conducted using only studies for which the calculation of ratios or odds ratios is possible.

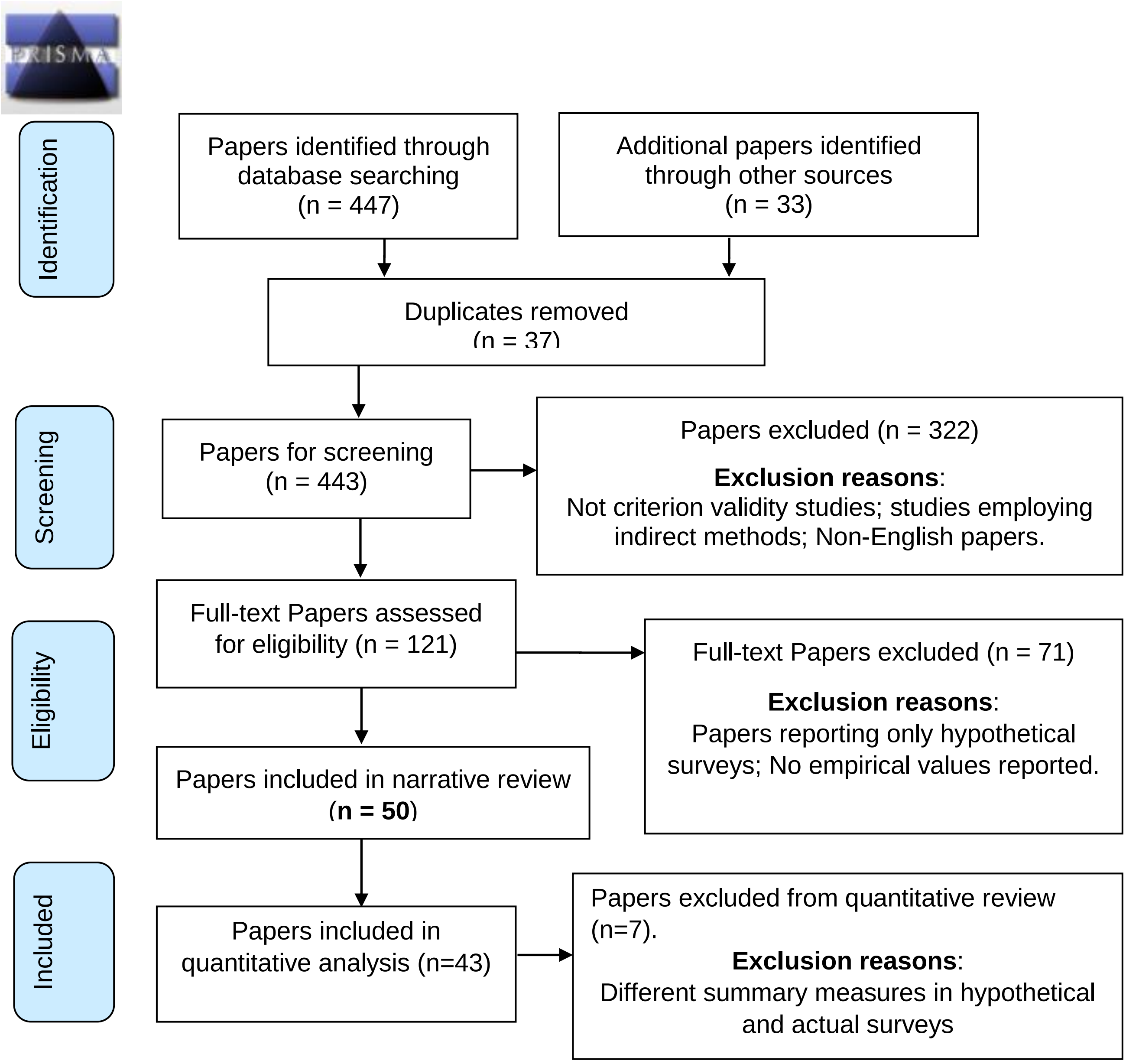
5.4.2 Results

5.4.2.1 Background characteristics

Of the 480 papers initially identified, 50 papers were included in the qualitative analysis and 43 for the quantitative analysis (figure 5.2). From the 50 papers, a total of 159 comparisons of hypothetical and actual values goods and services conducted across 14 countries were included in the reviews. 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 of hypothetical and actual values ranging from two to thirty. These included comparisons of the same good or service within the same study using different respondents or WTP elicitation methods. The results therefore, except for country and year of publication, focus on 159 comparisons of hypothetical and actual WTP (WTA) values from 50 papers. A summary of some of the background characteristics of all the papers included in the review is provided in Appendix 13.

The majority of comparisons (n=94), did not explicitly use any specific terms for validity assessment (such as defined in appendix 3), preferring to reflect papers as testing comparisons between hypothetical and actual WTP values. Approximately one fifth of these (n=32) referred to this as testing for hypothetical bias (Blumenschein et al. 1998; Botelho & Pinto 2002; Bryan & Jowett 2010; Camacho-Cuena et al. 2004; Getzner 2000; Johannesson 1997; Mozumder & Berrens 2007; Murphy et al., 2002; Onwujekwe et al. 2005). Two comparisons used the term predictive validity (Onwujekwe 2001), while one used external validity (Muller & Ruffieux 2011). Approximately one-fifth (n=30) of the comparisons referred to assessments of criterion validity (Bhatia & Fox-Rushby 2003; Bratt 2010; Carlson 2000; Johnston 2006; Loomis et al. 1996; Onwujekwe et al. 2001; Onwujekwe & Uzochukwu 2004; Onwujekwe 2004; Ramke et al. 2009; Vossler, Ethier, et al. 2003; Vossler & Kerkvliet 2003).

Figure 5-2: Flow of papers during the search process


PRISMA
Identification
Screening
Eligibility
Included
Papers identified through database searching (n = 447)
Additional papers identified through other sources (n = 33)
Duplicates removed (n = 37)
Papers for screening (n = 443)
Papers excluded (n = 322)
Exclusion reasons: Not criterion validity studies; studies employing indirect methods; Non-English papers.
Full-text Papers assessed for eligibility (n = 121)
Full-text Papers excluded (n = 71)
Exclusion reasons: Papers reporting only hypothetical surveys; No empirical values reported.
Papers included in narrative review (n = 50)
Papers included in quantitative analysis (n=43)
Papers excluded from quantitative review (n=7).
Exclusion reasons: Different summary measures in hypothetical and actual surveys

The comparisons were predominantly taken from the other (40%) and environmental (38%) and 22% from the health sector. The “other” sector includes comparisons of goods such as household and personal goods which do not fall under the environment or health sectors. More than half of the comparisons between hypothetical and actual WTP were for pure private goods (n=80), fifty-five were for public goods predominantly in the environment sector and 24 were for quasi-private goods (Table 5.1).

Table 5-1: Type of good valued by sector

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

Of the 36 health sector comparisons, 30 elicited values for prevention products such as treated mosquito nets (M R Bhatia & Fox-Rushby 2003) and six elicited values for management or treatment of a disease condition (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). From other sectors, while nearly all comparisons (n=54) elicited values for personal and household goods (e.g. art prints (Loomis et al. 1997; Loomis et al. 1996), sunglasses (Blumenschein et al. 1998)), one elicited values for both a personal good (chocolate bar) and a public good (prevention of additional damages to an aquatic system from acid rain) (Kealy & Dovidio 1990).

5.4.2.2 Comparison of hypothetical and actual survey attributes

All comparisons adopted a cross-sectional survey design for the survey of hypothetical and then actual WTP values. Nearly all the comparisons elicited WTP estimates (n=154) while WTA values were derived in 5. One, in the environment (Heberlein & 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 (Bishop & Heberlein 1986) . Four WTA comparisons were conducted in the other sector and these include eliciting expected compensation values from respondents in exchange for their holiday gifts followed by offers of actual payments for their holiday gifts (List & Shogren 2002).

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. 2003; Vossler & Watson 2013; Vossler & Kerkvliet 2003). In the same sector, two comparisons were asked for voluntary donations towards public good (Macmillan et al. 1999; Veisten & Navrud 2006). The following section sets out the similarities and variability between the two types of survey, for a range of design attributes.

Elicitation formats

Most of the comparisons use 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 5.2 where for every attribute; the similarities across the hypothetical and actual surveys are represented by the diagonal which is in bold.

The most common elicitation format used in both the hypothetical and actual surveys was dichotomous choice (n=66), followed by open ended questions (n=27), auctions (n=11) and bidding method (n=7). The same elicitation format was used in both the hypothetical and actual surveys in around 69% of the comparisons (n=111), excluding those clustered under “others” (n=9). 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 (M R Bhatia & Fox-Rushby 2003; Vernazza et al. 2015) in the health sector. 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). In another case the actual survey was adjusted to allow for higher maximum values following the hypothetical survey where a dichotomous choice question was asked in the hypothetical survey and then in the actual survey a dichotomous choice is presented followed by an open-ended question. WTP for health good/services has most commonly used different elicitation formats for the hypothetical and actual surveys (90%).

It is typically the environment (n=53) and other (n=39) sectors that have used the same elicitation formats for both the hypothetical and actual surveys. In the health sector, 23 comparisons from 5 studies use the same elicitation formats in both the hypothetical and actual surveys (Blumenschein et al. 2001; Onwujekwe & Uzochukwu 2004; Blumenschein et al. 2008; Ramke et al. 2009; Loomis et al. 2009).

Table 5-2: Comparison of study attributes in hypothetical and actual surveys

	Actual Survey						
	Elicitation format	*Auction*	*Bidding game*	*Dichotomous Choice*	*Open ended*	*Others**	*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*	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**

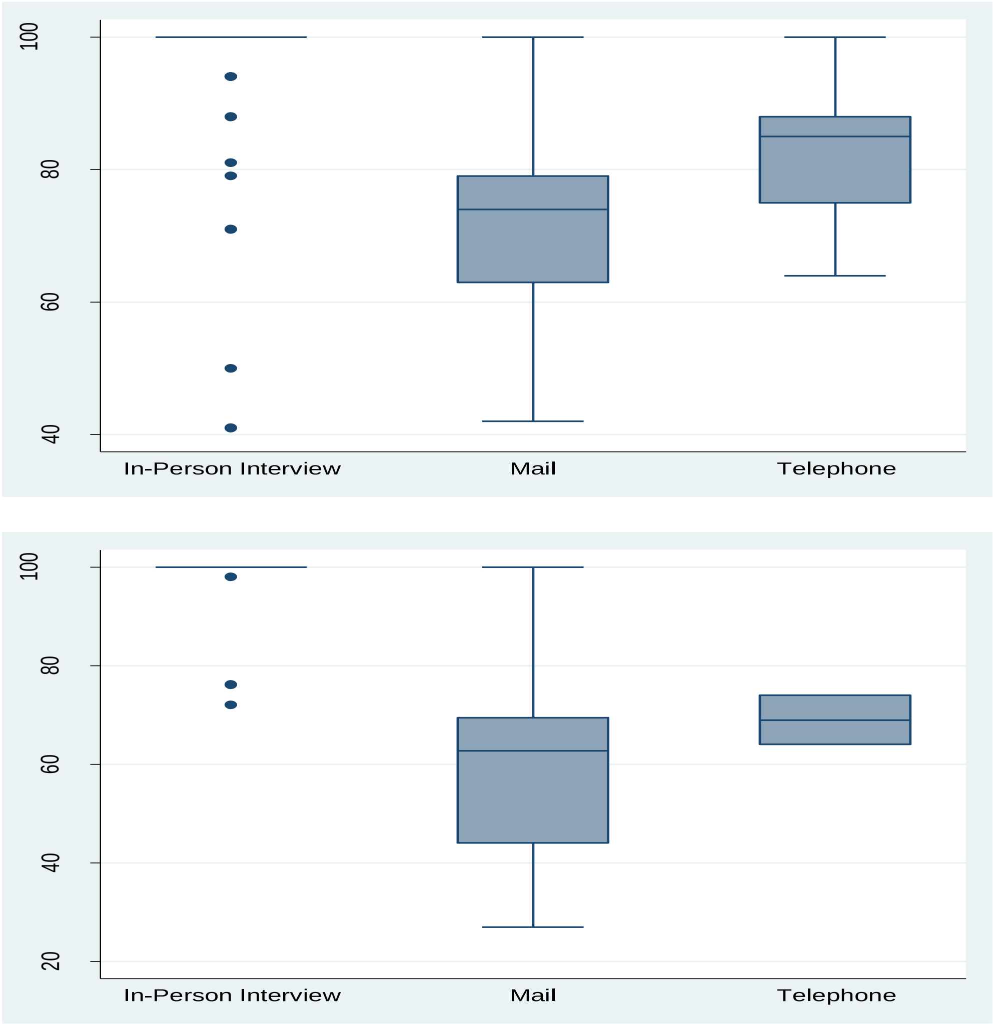
*****Other elicitation formats include all other elicitation formats with a count of less than 5 and these include structured haggling, payment cards and mixed methods such as binary or bidding game with follow up.

Survey administration

In-person interviews were used in 96 comparisons and postal surveys in 47. Different methods were used to administer the hypothetical and actual surveys in 16 comparisons. For example, a postal questionnaire was used for the hypothetical survey, but an interview was used in the actual survey in five comparisons; and interviews were used in the hypothetical survey in three comparisons while postal surveys were used in the actual surveys. The same mode of administration, interviews, was predominantly used in the “other” sector but different modes of administration were used in the health and environment sector. For example, in the health sector, one study utilised mail surveys in the hypothetical survey but in-person interviews in the actual survey (Loomis et al. 2009). In the environment sector, 4 comparisons used mail surveys for hypothetical values and interviews to elicit actual values (Vossler & Watson 2013; Vossler & Kerkvliet 2003; Johnston 2006) with three comparisons (2 studies) using the opposite (Brown & Taylor 2000; Seip Strand, J. 1992). In comparing the response rates by study modes of administration, in-person interviews and telephone interviews yielded the highest response rates in both the hypothetical and actual surveys. Mail surveys were least used and yielded the lowest response rates too.

Figure 5.3 shows that in both the hypothetical and actual surveys, telephone interviews had the highest response rates followed by mail. Response rates from in-person interviews are scattered across the scale implying missing response values or outliers in the data.

Figure 5-3: Response rates by survey administration modes



Response rates

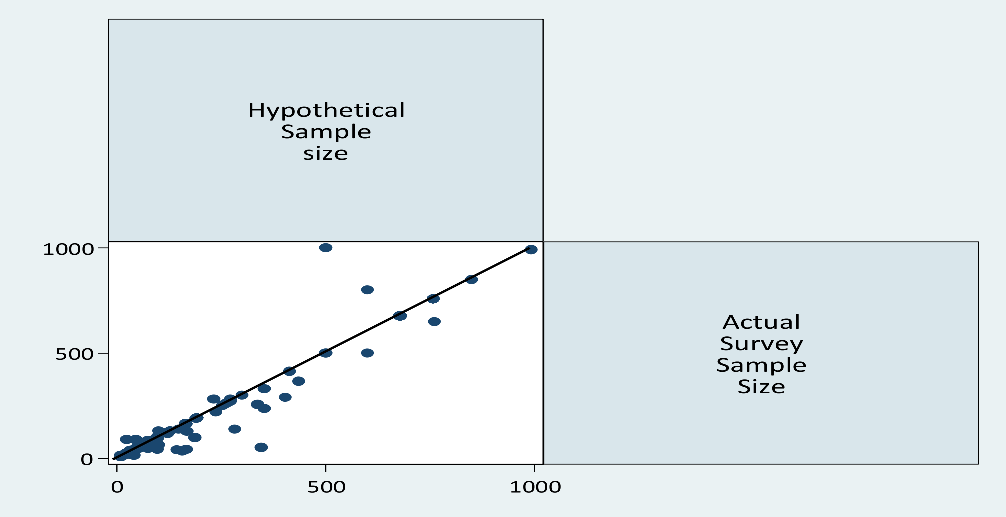
The general response rate is not indicated in more than two-thirds of the comparisons in the hypothetical surveys (n=63). The response rate for the WTP question specifically in the hypothetical survey is indicated in only one-third (n=53) of the surveys. In the actual survey, the general response rate is indicated in more than half the comparisons (n=130), while the response rate for the payment question is only indicated in 15 comparisons. In 13 cases (3 in health; 6 in environment; 4 in other sector), the response rate for the actual and hypothetical questions was the same.

Sample selection

Randomisation was used to select respondents in 27 comparisons (environment=8, health=13, “other” sector=6). Convenience samples, largely university students, were recruited for 54 comparisons. These were largely in the other (n=47) and environment sectors (n=7). Purposive samples which included potential users or beneficiaries of the goods or services being valued were used in 78 comparisons. These included asthma and diabetes patients in the health sector (Blumenschein et al. 2001; Blumenschein et al. 2008), museum attendees and goose hunters (Willis & Powe 1998; Bishop & Heberlein 1979; Heberlein & Bishop 1986) but most (n=45) were in the environment sector.

Figure 5.4 compares the sample sizes used in the hypothetical and actual surveys, with five outliers dropped from the summary (2 in hypothetical survey and 3 in the actual survey). Sample sizes ranged from 9 to 2,890 in the hypothetical survey and from 9 to 15,781 in the actual survey. The sample size for the two surveys was similar in 88 comparisons. In most cases, where the sample size differs, the hypothetical survey has a larger sample than the subsequent survey of actual values (n=44).

Figure 5-4: Samples sizes of surveys for hypothetical and actual CV values



Survey respondents

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. The same respondents completed the hypothetical and actual surveys in studies conducted across the mixed sector. 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 are not reported.

Duration between surveys

Hypothetical and actual surveys were undertaken concurrently or within a period of 2 weeks in most of comparisons (n=126), with 31 administering the two surveys within a period of more than 2 weeks apart while 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. 2003; Johnston 2006; Vossler & Watson 2013).

5.4.2.3 Justification for the values used in the surveys

For almost all elicitation formats, except open ended questions, pre-specified values are required for presentation to respondents in the survey e.g. a payment card presents a range of money values from which respondents are asked to select the value that best reflects their maximum WTP. 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, justification is not provided for value cues. In 7 comparisons from five studies, all in the environment sector, (Byrnes et al. 1999; Champ et al. 1997; Spencer et al. 1998; Champ & Bishop 2001; Blumenschein et al. 2008), the values presented to the respondents in both the hypothetical and actual surveys are based on prior costings of the planned projects. In another sample, (Loomis et al. 1997), values obtained from a pre-test of the survey are presented to respondents in both surveys.

Values from hypothetical surveys can be used to inform the actual survey, as in four comparisons (Bhatia & Fox-Rushby 2003; Loomis et al. 1997; Willis & Powe 1998; Onwujekwe 2004). In two comparisons, one each in the health and environment sector, the stated hypothetical values are presented in the actual survey (Onwujekwe 2004; Willis & Powe 1998). One study in the other sector (Loomis et al. 1997) presents the hypothetical mean value in all the comparisons with the hypothetical modal value presented in one study in the health sector (Bhatia & Fox-Rushby 2003). Market prices for the commodities are used in the actual surveys in one paper in the health sector (Onwujekwe et al. 2001). For 51 comparisons that used open ended questions, a justification is not relevant.

Most comparisons in the environment sector present the stated hypothetical values in the actual survey (n=34). In 11 comparisons the value presented in the actual survey is based on a costing of the proposed project. In the ‘other sectors’, nearly one-third of the comparisons (n=14) do not provide a justification for the values used in the hypothetical surveys, while the market price for the good is presented in the two comparisons in one study and the value is centred around the pre-test mean in two comparisons in another (Loomis et al. 1997). Auctions and open ended elicitation formats are used in the actual surveys in thirteen comparisons.

5.4.2.4 Valuation estimates

Table 5.3 shows that 84 comparisons presented summary means for both surveys while 60 provided summary percentages. Different summary estimates were provided for 15 sample pairs.

Table 5-3: Valuation estimates summary format in hypothetical and actual surveys

Hypothetical survey		Actual Survey		
		Mean	Percentage	Total
	Mean	**84**	13	97
	Percentage	2	**60**	62
	Total	86	73	**159**

5.4.2.5 Calculating WTP/WTA and testing criterion validity

The estimation methods for mean WTP are varied and whilst this would be expected to relate to question format, there are some cases where the method differs between the hypothetical and actual survey even though the same elicitation format is used (n=9).

The methods used to estimate WTP are not indicated in 25 hypothetical and 26 actual surveys. Along with this tends to be a lack of information on whether statistical tests of mean differences between hypothetical and actual WTP were conducted and, if so, how. Some simply compared the mean values and where similar concluded that the comparisons demonstrated validity and, in such comparisons, no indication of the spread of data was given. Roughly 70% (n=111) of the comparisons indicate the statistical tests used with the majority (n=88) employing parametric methods (non-parametric methods = 18, both parametric and non-parametric methods n=20).

Criterion validity was confirmed by authors in 17 comparisons presenting mean summaries in both surveys. The majority of confirmations were in comparisons from the other sector (n=15). Criterion validity was not confirmed by the authors in any of the health sector comparisons with only two confirmations from the environment comparisons. For studies reporting percentage summaries, criterion validity was confirmed by authors in 12 comparisons, majority of which are in the health sector (n=9) followed by the other sector (n=2). Only one environmental sector comparison confirmed criterion validity.

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 to 11.8). The greatest differences were typically for environmental goods/services. For example, in one study which elicited WTP for the protection of a sensitive rainforest land, the hypothetical mean WTP was $27.97 for female respondents and $72.22 for male respondents whereas the mean actual WTP was $3.23 among females and $6.14 for males (Brown & Taylor 2000). For the comparisons which presented percentage summaries in both hypothetical and actual surveys odds ratios were calculated for comparisons which had non-zero values in both surveys (n=56). The average odds ratio was 5.7 with a range 0 to 13.6. The ratios and odds ratios for the included comparisons by different design attributes are presented in Table 5.4 (overall study characteristics) and Table 5.5 (hypothetical and actual surveys).

For both mean and percent summaries, the highest ratios were found in valuation goods in the environment sector (ratio 5.99; OR 8.10), with within sample comparisons (ratio 3.27; OR 6.51), when hypothetical and actual surveys are conducted concurrently (ratio 3.24; OR 8.41) and when one-off payments are elicited (ratio 3.22; OR 6.01). Separately, for comparisons presenting mean summaries, the highest ratios were found in valuations with pure public (4.92), and conservation goods (5.96), and when a donation mechanism was used as the payment vehicle (4.53). Among the comparisons which presented percent summaries, the highest odds ratios were observed with quasi private goods (28.12), goods used for “other” purposes (17.77), for studies conducted within a laboratory setting (7.14) and when cash fee payments are elicited (6.77).

Table 5-4: Summary estimates by study attribute (Overall study characteristics)

Variable	Ratio [**SD**] *(no. of comparisons)*	Odds ratio [**SD**] (no. of comparisons)
Country income level		
i. High income	3.17 [**3.70**] *(84)*	7.35 [**17.15**] *(41)*
ii. Lower middle income	-	1.25 [**0.46**] *(15)*
Sector		
i. Health	1.75 [**0.70**] *(9)*	1.49 [**0.84**] *(15)*
ii. Environment	5.99 [**3.75**] *(23)*	8.10 [**19.95**] *(30)*
iii. Other [a]	2.17 [**3.35**] *(52)*	4.98 [**4.19**] *(11)*
Class of good or service		
i. Pure Public	4.92 [**3.68**] *(22)*	5.02 [**4.800**] *(26)*
ii. Pure Private	2.49 [**3.67**] *(42)*	2.97 [**3.24**] *(26)*
iii. Quasi-Private	2.70 [**3.31**] *(20)*	28.12 [**55.13**] *(4)*
Purpose of good or service		
i. Prevention	1.75 [**0.70**] *(9)*	1.56 [**0.80**] *(19)*
ii. Conservation	5.96 [**3.78**] *(23)*	4.67 [**4.80**] *(28)*
iii. Other [c]	2.19 [**3.34**] *(52)*	17.77 [**35.13**] *(9)*
Type of comparison		
i. Between	3.12 [**4.06**] *(53)*	2.49 [**2.89**] *(11)*
ii. Within	3.27 [**3.04**] *(31)*	6.51 [**16.47**] *(45)*
Survey setting		
i. Field	0.98 [**0.83**] *(29)*	5.55 [**15.68**] *(50)*
ii. Laboratory	0.66 [**0.75**] *(55)*	7.14 [**4.69**] *(6)*
Duration between surveys		
i. Concurrent	3.24 [**3.86**] *(76)*	8.41 [**21.57**] *(25)*
ii. 1-7 days	2.78 [**1.34**] *(7)*	6.94 [**6.13**] *(12)*
iii. More than 7 days	1.08 [**-**] *(1)*	1.41 [**1.40**] *(19)*
Payment Vehicle		
i. Cash Fee	2.83 [**3.69**] *(67)*	6.77 [**20.25**] *(29)*
ii. Donation	4.53 [**3.51**] *(17)*	5.12 [**4.92**] *(24)*
iii. Property tax	-	0.34 [**0.58**] *(3)*
Payment duration		
i. Annual Payment	1.39 [**0.44**] *(2)*	3.86 [**4.03**] *(2)*
ii. One-Off	3.22 [**3.73**] *(82)*	6.01 [**15.39**] *(52)*
iii. Monthly	-	0.003 [**0.003**] *(2)*
Money effects		
i. Money given for participation in surveys	1.39 [**0.44**] *(2)*	1.39 [**0.44**] *(2)*
ii. Money given for purchase of good in actual survey	1.39 [**0.44**] *(2)*	1.39 [**0.44**] *(2)*
Overall	3.17 [**3.70**] *(84)*	5.72 [**14.87**] *(56)*

[a] Includes consumer goods such as books, sunglasses. [c] Includes consumables such as food, clothing and household items

Table 5.5 presents the ratios and odds ratios for a select range of hypothetical and survey design attributes. For both hypothetical and actual surveys, the highest ratios were noted with non-students and mixed student samples (10.21 in both), non-users (4.75 in both) and the use of open ended WTP elicitation methods (hypothetical 5.10; actual 5.49). For the mean summaries, in the hypothetical surveys, the highest ratios were found with purposive samples (3.60) and telephone interviews (5.60). In the actual surveys, the highest ratios were found when random surveys were used (4.22), with mail surveys (5.05). Where odds ratios are presented, these are similar in both hypothetical and actual surveys for both. The highest odds ratios are observed in both the hypothetical and actual surveys when purposive samples are used (7.93 and 7.15 respectively); with student samples (7.14 in both); where the respondents are potential users of the valuation good (5.89 and 5.72 respectively); when in-person interviews are used (6.77 and 6.57 respectively) and when open ended survey elicitation formats are used (17.86 and 7.53 respectively).

To further the analysis, a comparison was made of ratios and odds ratios for similarities (or differences) in key design attributes in both the hypothetical and actual surveys of WTP. The results of this comparison are presented in table 5.6. For both mean and percent summaries, the ratios (and odds ratios) are highest in comparisons of hypothetical and actual surveys of WTP where for both; different elicitation methods (ratio 3.77; OR 28.59) and specifically open-ended methods (ratio 6.13; OR 7.53) are used. Where mean summaries are presented, ratios are higher when for both hypothetical and actual surveys student samples (ratio 3.34) and non-users of the valuation good are used (4.75), different sample selection methods (4.33) and particularly with the use of convenience samples in both surveys (3.13). The ratios are also highest when for both hypothetical and actual surveys different survey administration modes (6.27), particularly mail administration are used (4.03).
With percent summaries, the highest odds ratios are found with the use of non-student samples (7.14), respondents who are users of the valuation good (5.89), the same sample selection method (6.07), and particularly with the use of purposive samples in both hypothetical and actual surveys (7.93). The odds ratios are also high when the same administration mode is used in comparisons hypothetical and actual surveys (6.01), particularly with the use of in-person surveys (6.77).

Table 5-5: Summary estimates by key design attributes (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)*	1.66 [**1.50**] *(19)*	1.80 [**1.67**] *(15)*
ii. Purposive	3.60 [**3.25**] *(32)*	2.86 [**2.73**] *(31)*	7.93 [**19.62**] *(31)*	7.15 [**18.56**] *(35)*
iii. Convenience	3.05 [**4.10**] *(48)*	3.13 [**4.37**] *(42)*	7.14 [**4.69**] *(6)*	7.14 [**4.69**] *(6)*
Sample type (1)				
i. Students	2.98 [**4.54**] *(33)*	2.94 [**4.48**] *(34)*	7.14 [**4.69**] *(6)*	7.14 [**4.69**] *(6)*
ii. Non-Students	3.02 [**2.77**] *(49)*	3.05 [**2.80**] *(48)*	5.55 [**15.68**] *(50)*	5.55 [**15.68**] *(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)*	5.89 [**15.12**] *(54)*	5.72 [**14.87**] *(56)*
ii. Non-Users	4.75 [**3.84**] *(9)*	4.75 [**3.84**] *(9)*	0.97 [**0.05**] *(2)*	-
Study administration mode				
i. In-person interviews	2.78 [**3.71**] *(60)*	2.47 [**3.40**] *(61)*	6.77 [**20.25**] *(29)*	6.57 [**20.28**] *(29)*
ii. Mail Surveys	3.88 [**3.38**] *(20)*	5.05 [**3.87**] *(23)*	4.59 [**4.87**] *(27)*	5.19 [**4.82]**_(25)_
iii. Telephone	5.60 [**4.66**] *(4)*	-	-	0.003 [**0.0003**] *(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)*	1.10 [**0.17**] *(7)*	1.11 [**0.18**] *(5)*
iii. Dichotomous choice	3.08 [**2.54**] *(29)*	2.18 [**1.72**] *(34)*	3.55 [**3.30**] *(37)*	5.99 [**17.08**] *(41)*
iv. Open ended	5.10 [**5.40**] *(27)*	5.49 [**4.14**] *(17)*	17.86 [**33.24**] *(10)*	7.53 [**6.57**] *(9)*
v. Payment Card	1.17 [**0.42**] *(7)*	3.05 [**3.75**] *(9)*	-	-
vi. Referendum	1.15 [**0.24**] *(8)*	-	-	-
vii. Binary with follow up	-	-	1.29 [**0.15**] *(2)*	1.18 [**-**] *(1)*

Table 5-6: Comparisons of key study attributes in hypothetical and actual surveys

Variable name	**Value labels**	**Mean Summaries**	**Percent Summaries**
		Ratio [**SD**] (*no. of comparisons*)	Odds Ratio [**SD**] (*no. of comparisons*)
Student sample	Students in both surveys	3.34 [**4.68**] *(36)*	5.55 [**15.68**] *(50)*
	Non-students	3.05 [**2.80**] *(48)*	7.14 [**4.69**] *(6)*
Users	Respondents users / potential users of valuation good	2.99 [**3.66**] *(75)*	5.89 [**15.12**] *(54)*
	Non-users	4.75 [**3.84**] *(9)*	0.97 [**0.05**] *(2)*
Sample selection	Same method in both surveys	2.96 [**3.81**] *(71)*	6.07 [**15.39**] *(52)*
	Different methods	4.33 [**2.87**] *(13)*	1.14[**0.20**] *(4)*
Sample selection categories	Random sampling in both surveys	1.27 [**0.37**] *(4)*	1.80 [**1.67**] *(15)*
	Convenience sampling in both surveys	3.13 [**4.37**] *(42)*	7.14 [**4.69**] *(6)*
	Purposive sampling in both	2.95 [**3.03**] *(25)*	7.93 [**19.62**] *(31)*
Money effects (1)	Money given in both surveys	2.81 [**3.35**] *(34)*	5.06 [**4.77**] *(34)*
Money effects (2)	Respondents given cash in actual survey for purchase	2.81 [**3.35**] *(34)*	6.09 [**5.42**] *(16)*
Administration mode	Same mode on both surveys	2.89 [**3.50**] (*77*)	6.01 [**15.39**] *(52)*
	Different modes	6.27 [**4.64**] *(7)*	1.93 [**3.22**] *(4)*
Administration mode categories	Mail administration in both surveys	4.03 [**3.41**] *(19)*	5.12 [**4.92**] *(24)*
	In-person surveys in both surveys	2.52 [**3.48**] *(58)*	6.77 [**20.63**] *(28)*
Elicitation method	Same method in both surveys	3.09 [**2.99**] *(58)*	3.96 [**4.21**] *(52)*
	Different methods	3.77 [**5.00**] *(26)*	28.59 [**54.81**] *(4)*
Elicitation mode categories	Auction methods	3.05 [**4.88**] *(23)*	-
	Bidding methods	1.60 [**-**] *(1)*	1.11 [**0.18**] *(5)*
	Dichotomous methods	2.53 [**1.89**] *(25)*	3.49 [**3.27**] *(38)*
	Open ended methods	6.13 [**4.25**] *(14)*	7.53 [**6.57**] *(9)*

5.5 Discussion

This review shows that a 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. With the increasing use of simulated market experiments, it is not surprising that the majority 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 2/3 of investigations are from US, with the remaining 35% spread across 9 countries. The evidence therefore covers a wider range of applications, with the majority focussed on the US for private goods. 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 variety in language used to describe the objectives of research in this area reflects much variety in thought and some confusion. Not only does the definition of external/criterion validity differ 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 the evidence base was accessed through reviews of references and citation searching. Future reviews might therefore consider a wider variety of search terms but expect this to be costly in the very large numbers of titles and abstracts returned for review.

This review gives the current indication of the degree of variation in stated and actual WTP in the CV literature; hypothetical WTP (WTA) was on average 3.2 times greater (lower) than actual WTP (WTA), with a range of 0.7-11.8 for mean summaries and 5.7 (range, 0-13.6) for percent summaries. It also shows that current conclusions are heavily weighted (82% agreement) towards claims that criterion validity is not demonstrated, as only 18% authors claim evidence of criterion validity. However, alongside this evidence, there is neither discussion of ‘how close is close enough?’ nor consideration of how valid the presented evidence itself was and therefore I question whether the results are quite as robust as they appear to be.

This review has also shown 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 45% 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 in fact be different.

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 that can be used 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. Majority of the reviewed comparisons do 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 (Alwin 1992). However, while longer durations could potentially introduce recall bias, short durations of the time difference mean that the respondent may remember what they said in the hypothetical survey and deliberately repeat the value to appear publicly consistent. Further, while a longer duration between the two surveys might offer the respondent sufficient time to think and time to forget their original values, it also increases the possibility of real change occurring and 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. The effect of this has not yet been tested.

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 seem to use some appropriate methods compared with other sectors. For example, higher proportions of comparisons use the same respondents, administration modes, elicitation formats and payment vehicles in hypothetical and actual surveys. Comparisons also report on key explanatory variables, allowing a comparison within the sector and potential transferability of the methods used to assess criterion validity across sectors. Having the same respondent responding to the hypothetical and actual valuation scenarios reduces bias when judging criterion validity and this too occurs 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 are 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.

The review also indicates 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, 1/3rdof 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. We noticed too how few explanations were given for the selection of value cues behind bid offers regardless of design format. Until such time as 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. 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 contingent valuation comparisons, as poor reporting continues to limit the use of comparisons for systematic reviews and meta analyses in contingent valuation research (Trapero-Bertran et al. 2013). Guidelines on the conduct and reporting of criterion validity assessments are also needed if CV-WTP studies are to be broadly adopted in benefit assessments.

The confusion in the terms used to describe validity might have led to some papers being missed from the search. However, the reference lists for all the papers identified from the database search were screened for additional papers. An author search was also conducted in google to identify more papers. I believe that the search was exhaustive. Based on this process, and in the absence of guidelines on the conduct of criterion validity assessments, robust systematic reviews on criterion validity will be significantly informed more by the use of methods other than conventional database searches.

Further, even with the robust methods used to identify empirical studies for this review, relatively few criterion validity assessments were found. This limits robust analysis and interpretation of the results. However, the reviews utilised all the estimates reported in each of the studies, which significantly increased the dataset. Even then, significantly fewer studies were identified from the health sector. Ultimately, this limits the generalisability of the conclusions on hypothetical bias to the health sector. However, the findings and conclusions on the methods are generalizable across the sectors.

5.6 Conclusions and chapter summary

The current evidence on criterion validity has been summarised in this chapter. In the first section, a summary of reviews on criterion validity was presented. This was followed by a systematic review of primary studies assessing the criterion validity of

CV-WTP methods. The evidence on the criterion validity for contingent valuation comparisons is more mixed than authors are representing. Evidently, substantial differences exist in study design between hypothetical and actual WTP/WTA surveys. These differences are not accounted for and could be the reason for the reported hypothetical bias. This concern is compounded by the presence of key gaps in the reporting of methods and data.

The WTP method offers potential for 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. Infact, some authors have acknowledged that poor criterion validity conclusions may have been made based on poorly conducted studies that incorrectly suggest a lack of criterion validity. Significant differences in the reporting of and analysis of CV studies is noted. Interestingly, some of the earlier reviews are based on re-analysis of existing reviews with the authors citing errors in previous analysis. In one of the reviews, the authors “correct” data extraction errors in an earlier review. These efforts are acknowledged in the ongoing attempts to improve the method. The development of reporting guidelines for contingent valuation comparisons and the development of methodological guidelines for the conduct of criterion validity assessments would aid assessment of validity and transferability of results.

With the last review of criterion validity assessments conducted more than one decade ago, the systematic review presented in this chapter presents the current evidence on the criterion validity of CV-WTP studies. While the growth in conduct of CV studies is noted, there are still very few studies conducted in the health sector. The limited assessments in the sector pose a challenge to investigations of criterion validity which might inform the adoption of the CV method. Using estimates from the studies identified in the systematic review, the magnitude of hypothetical bias is quantified in the next chapter. As with the review, this presents the current evidence on the magnitude of hypothetical bias in CV-WTP studies. Meta-regressions are also conducted to explore the drivers of hypothetical bias. The results of these analyses will support the development of an explanatory model of variation of hypothetical and actual WTP values.

Chapter 6 A Meta-analysis and Meta-regression of Studies Assessing the Criterion Validity of WTP Methods

The systematic review discussed in the previous chapter identified methodological differences in the design and conduct of hypothetical and actual surveys of WTP used to assess criterion validity. In this chapter, the WTP values presented in the reviewed papers are meta-analysed to generate a pooled estimate of the magnitude of hypothetical bias. With the high level of heterogeneity in the studies, robust meta-regressions are conducted separately for studies presenting mean and percent summaries to determine the factors which are associated with hypothetical bias in CV WTP studies. In the introduction of this chapter, a brief overview of hypothetical bias is discussed, including the current evidence on the estimated magnitude. The methods used in the meta-analysis and the meta-regression are then discussed with the results presented after. The analyses presented in this chapter provide an update from the last similar analyses conducted more than one decade ago on the magnitude of hypothetical bias in CV-WTP studies.

6.1 Introduction

As discussed in previous chapters, the criterion validity of CV WTP has been the subject of criticism among various authors over time. Assessments of criterion validity involve the estimation of the divergence between hypothetical WTP statements and actual values. This divergence is referred to as hypothetical bias. In the last published meta-analysis of hypothetical bias in stated preference valuation, Murphy et al (2005) estimated hypothetical bias to be 2.6 (median 1.35). This meta-analysis builds on an earlier systematic review of studies investigating hypothetical bias (List & Gallet 2001) in which the range of hypothetical bias in 29 identified studies ranged from 0.80 – 1.50. Empirical evidence by Harrison and Rutstrom (1999) empirical evidence suggests hypothetical bias of 338 percent while List and Gallet (2001) report hypothetical bias of a factor of three.

Early efforts to adjust hypothetical values to enable their use in policy include the proposal by the National Oceanic and Atmospheric Administration (NOAA) panel of 1993 to divide hypothetical values by a factor of 2 (Arrow et al. 1993). While the choice of this calibration factor has been questioned by different authors (Carson et al., 1996; Harrison &Rutström 2008; Murphy et al. 2005), Murphy (2005) in the most recent meta-analysis concludes that hypothetical bias in stated preference studies may not be as important as most previous studies suggest. In addition to assessing the magnitude of the bias present in stated preference studies, authors attempt to identify the factors responsible for this bias, with differing results. In the absence of a theory to explain hypothetical bias, these explorations have largely been hypothesis generating.

The current meta-analysis and meta-regression builds on the existing evidence by using an expanded dataset of criterion validity assessments published up to January 2017. In particular, this analysis includes more studies conducted in the health sector and a higher number of estimates allowing for more robust analyses to be conducted. The analysis differs from the previous meta-analyses in three ways. First, it includes only studies which use direct stated preference techniques to elicit WTP values. Second, only those studies which conduct both hypothetical and actual surveys, with an actual transaction are included. Third, it includes only those studies which provide empirical estimates of WTP/WTA, I do not attempt to calculate or model these from the provided data. In the next section, the methods used in the meta-analysis and meta-regression are provided.

6.2 Methods

As with the narrative review presented in chapter 5, results of the analyses are reported by comparisons rather than by paper or study. This is to allow for the use of all the estimates provided in the papers and hence a larger dataset for the analysis. For all comparisons, WTP estimates for hypothetical and actual data are matched as pairs, when provided, and compared as a ratio (for mean values) and as odds ratios (for percentage summaries). The final dataset for this quantitative analysis includes ratios for 84 comparisons and odds ratios for 60 comparisons. The natural logs of the ratios and the odds ratios are used in the analysis. All the analyses are conducted separately for comparisons presenting mean and percent summaries.

Matching of pairs was not possible for 15 comparisons and these are excluded in the analysis. Thirteen of these presented mean summaries in the hypothetical surveys and percentages in the actual surveys while 2 had percentages in the hypothetical surveys and mean summaries in the actual surveys. The background characteristics of all the papers (n=43) included in the analyses are detailed in Appendix 14. All quantitative analyses were conducted using metan (Bradburn et al. 1998) and metareg (Sharp S 1998) commands in Stata14 and these are further discussed in the following sections.

6.2.1 Meta-analysis

An approach developed by DerSimonian and Laird (1986) was used to perform a random-effects meta-analysis, summarizing the log ratios and odds ratios. This model assumes that the observed values are a random sample from a distribution of values with equal variance (Harris et al. 2008). The comparisons are weighted by the inverse of the variance of the effect estimate. For the mean summaries, only comparisons which report standard errors of the mean, or those which provide sufficient information to enable the calculation of the standard error are included. Only comparisons which have non-zero hypothetical and actual WTP value (and hence a non-zero log odds ratio) are included in the meta-analysis for percentage summaries. Forest plots are generated separately for the mean and percentage summaries and the I^2 is used to determine the level of heterogeneity (Higgins et al., 2003).

To further explore the sources of the heterogeneity, sub-group analysis were conducted in an exploratory manner for a range of characteristics such as sector, sample type (student versus non-student sample), payment vehicle (taxes, donations, out of pocket payment), respondent type (within or between sample comparisons) and survey setting (field or laboratory). In the sensitivity analysis, meta-analyses are re-run excluding comparisons with the highest samples.

6.2.2 Meta-regressions

Meta-regressions are conducted to explain the heterogeneity in the presented summaries and determine the drivers of hypothetical bias. These regressions are run separately for studies reporting mean and percent summaries. All of these are clustered by study to control for the multiple comparisons from some of the studies. All the comparisons reporting mean summaries and only those comparisons which report a non-zero odds ratio among the percentage summaries are included in the regression models.

The dependent variables in these regressions are: (1) the log ratio of hypothetical to actual values for comparisons presenting mean summaries and; (2) the log of the odds ratio of actual to hypothetical values for comparisons presenting summaries as percentages. Previous meta-regressions have investigated the effect of different study attributes on hypothetical bias (Carson et al. 1996; Harrison & Rutström 2008; Liljas & Blumenschein 2000; List & Gallet 2001; Little & 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 are 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 and non-users); (7) WTP elicitation format in both surveys; (8) type of comparison (either between or within); (9) study setting (laboratory or field); (10) duration between the hypothetical and actual surveys and (11) found money effects (whether respondents are paid to participate in either survey or given money to purchase the valuation good).

For all the regressions, the independent variables are entered the models as binary variables. Correlation coefficients were determined for all independent variables that were selected for inclusion in the regression models (appendix 31). The Spearman's correlation coefficient, regarded as more robust to outliers than the Pearson's correlation coefficient was used (Mukaka 2012). The rule of thumb of interpreting the size of a correlation coefficient as suggested by Hinkle (Hinkle et al. 2003) was applied (see appendix 32). Where variables presented with moderate to very high positive (negative) correlation coefficients (>0.50) a pragmatic decision was made to determine the choice of variable to include in the regression model.

In additional analyses, the variables sector, class and purpose of the good are entered as dummy variables representing the respective categories. All the variables included in the analysis are defined in Appendix 15.

6.2.2.1 Univariate meta-regressions

The range of univariate regressions explores the relationship of difference between the dependent variable and independent variables listed in the previous section separately for comparisons presenting mean and percentage summaries. The model outputs are presented and discussed in the results section.

6.2.2.2 Multiple meta-regression

Where the ratio is the dependent variable (comparisons presenting mean summaries), the GLM estimator is used. The GLM permits the use of the estimates in their natural form, with a straight forward interpretation. Where the odds ratio is presented, the natural log is used, and a logit model estimated. Base and reduced models are determined separately for comparisons summarized as means and percentages. In the base models, all the independent variables summarized in section 6.2.2 (and appendix 15) are included.

Standard approaches were used to address potential over-specification of the models. First, a parsimonious model was fitted for the regression model using a stepwise variable selection approach (Zhang 2016). Model diagnostics were undertaken using the link test with every estimation[22]. In all cases, models were found to be well specified as relevant parameter (haqstat) from this analysis was not significant. The final reduced model includes the range of variables which significant and for which the model is best specified. Finally, for each of the models, a predicted ratio or log odds ratio is determined for the mean and percent summaries respectively.

Where the log ratio is the dependent variable (for comparisons presenting mean summaries), the equation below was fitted allowing for heteroscedasticity tests:

$\log yi = \beta_o + \beta_1 x_{1i} + u_i + v_i x_{1i} + \cdots \beta_n x_n + u_n + v_n x_{ni}$where;

Log y_1=the log ratio of actual values to hypothetical values; β_o= constant; $x_i \ldots . x_n$ = predictor variables; u_1= residual associated with the intercept β_o; and v_1 = the residual associated with the slope parameter β_o of x_1.

[22] The Linktest is used to check for the goodness of fit of the models (Cameron & Trivedi, 2009). When a model is correctly fitted, the haqstat estimate is not significant. The Hosmer Lemeshow test was also used to check for the goodness of fit of the models (Archer & Lemeshow 2006; Hosmer & Lemeshow 2013)

With the log odds ratio as the dependent variable (for comparisons presenting percentage summaries), the following logit function which takes into account the effect of a predictor value on another predictor variable was fitted into the model:

$$logit\ (p) = \log\left(\frac{p}{1-p}\right) = \beta_o + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

For both the mean and percent summaries, base models are run with all the independent predictors, weighted by the study. The base models are compared with the reduced model obtained from the stepwise selection models in the results section.

6.3 Results

The results of the different analyses are presented separately for mean and percent summaries. The meta-analysis results are presented first, and these are followed by the meta-regression results.

6.3.1 Meta-analysis results

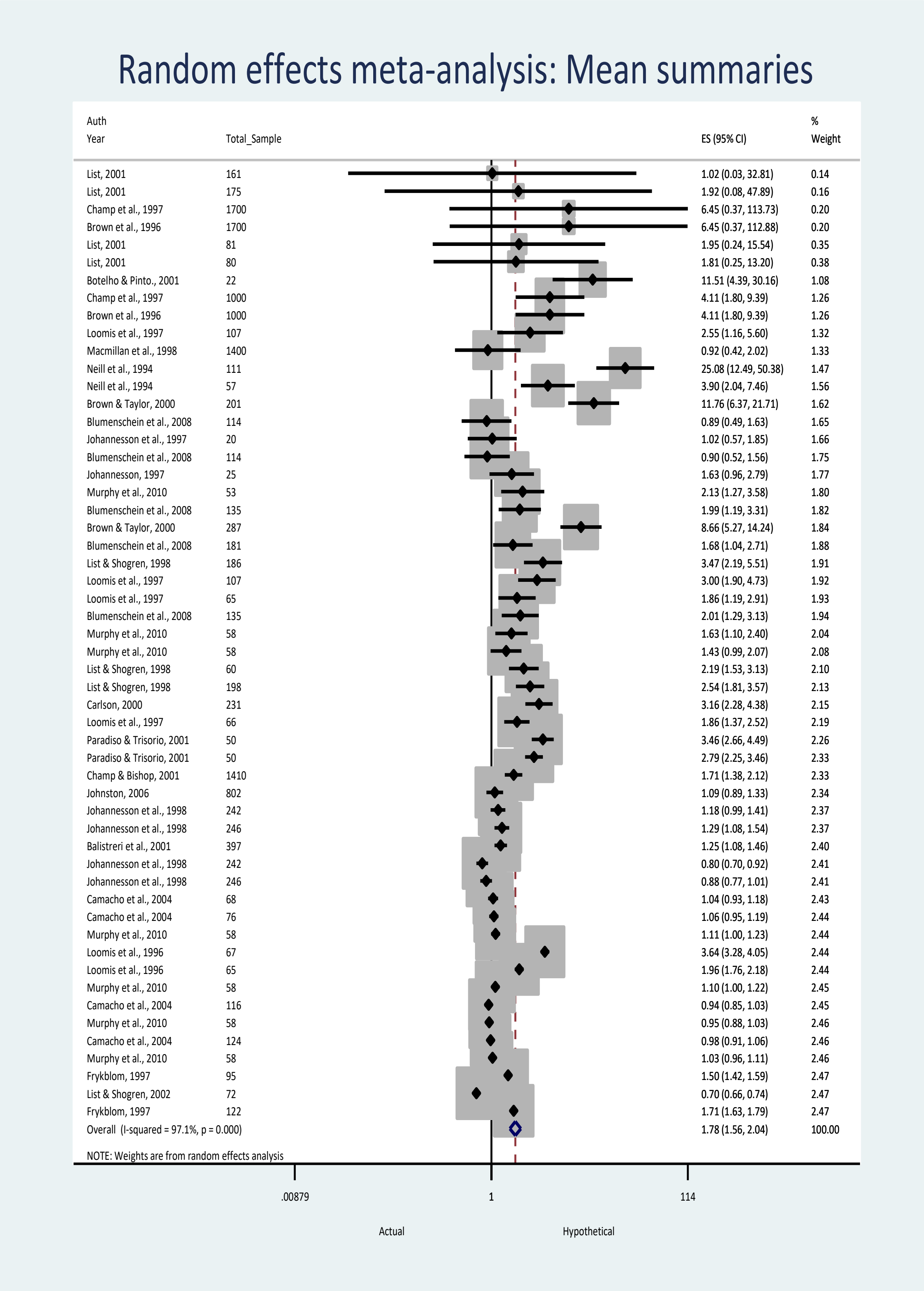
Meta-analysis was conducted separately for comparisons presenting mean and percentage summaries. Standard errors were provided or calculated where possible for a total of fifty-four of the comparisons presenting mean summaries and only these were included in the meta-analysis. For comparisons presenting percentage summaries, four reported a zero value in the actual survey results, generating an odds ratio of zero. These were excluded from the main analysis thus leaving a total of 56 comparisons. In a sensitivity analysis, zero WTP values are replaced with 0.001 and 0.0001 and the analyses re-run for comparison. Fifteen comparisons which presented different summaries in the hypothetical and actual surveys were excluded from the meta-analysis.

6.3.1.1 Comparisons presenting mean summaries

The ratio of the actual and hypothetical mean values was used in the random effects meta-analysis. From the forest plot (figure 6.1), almost all the lines fall on the right-hand side of the graph. This implies that hypothetical mean WTP overestimated actual mean values by a ratio of more than 1. The overall effect size for the comparisons presenting mean summaries is 1.785 and the variation in the ratio of mean hypothetical and actual values ranged from 1.562 to 2.038. The level of variation in the effect size attributable to heterogeneity in the comparisons included in this meta-analysis is indicated by the $I^2$of 97.1% and this is statistically significant (p=0.000).

In the subgroup analysis, differences in the level of heterogeneity were observed across the sector in which the study was conducted, and the survey setting. While the number of studies from this sector are few, the level of heterogeneity in the health sector was significantly lower with an I^2 of 56.5% (p=0.056). The levels of heterogeneity in the “Other” and Environment sectors were 97.7% and 92.7% respectively with p values of 0.000 in both (figure 6.2). In the subgroup analysis by survey setting, while the overall level of heterogeneity remained high and significant regardless of study setting (97.1%, p=0.000), this was much lower with field studies (68.4%, p=0.000) compared to laboratory studies (97.7%, p=0.000) (see figure 6.3). In a sensitivity analyses, the effect on the pooled ratio of dropping comparisons which had the widest confidence intervals were 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=0.000) (Figure 6.4).

Figure 6-1: Forest plot of log ratios for comparisons reporting mean summaries

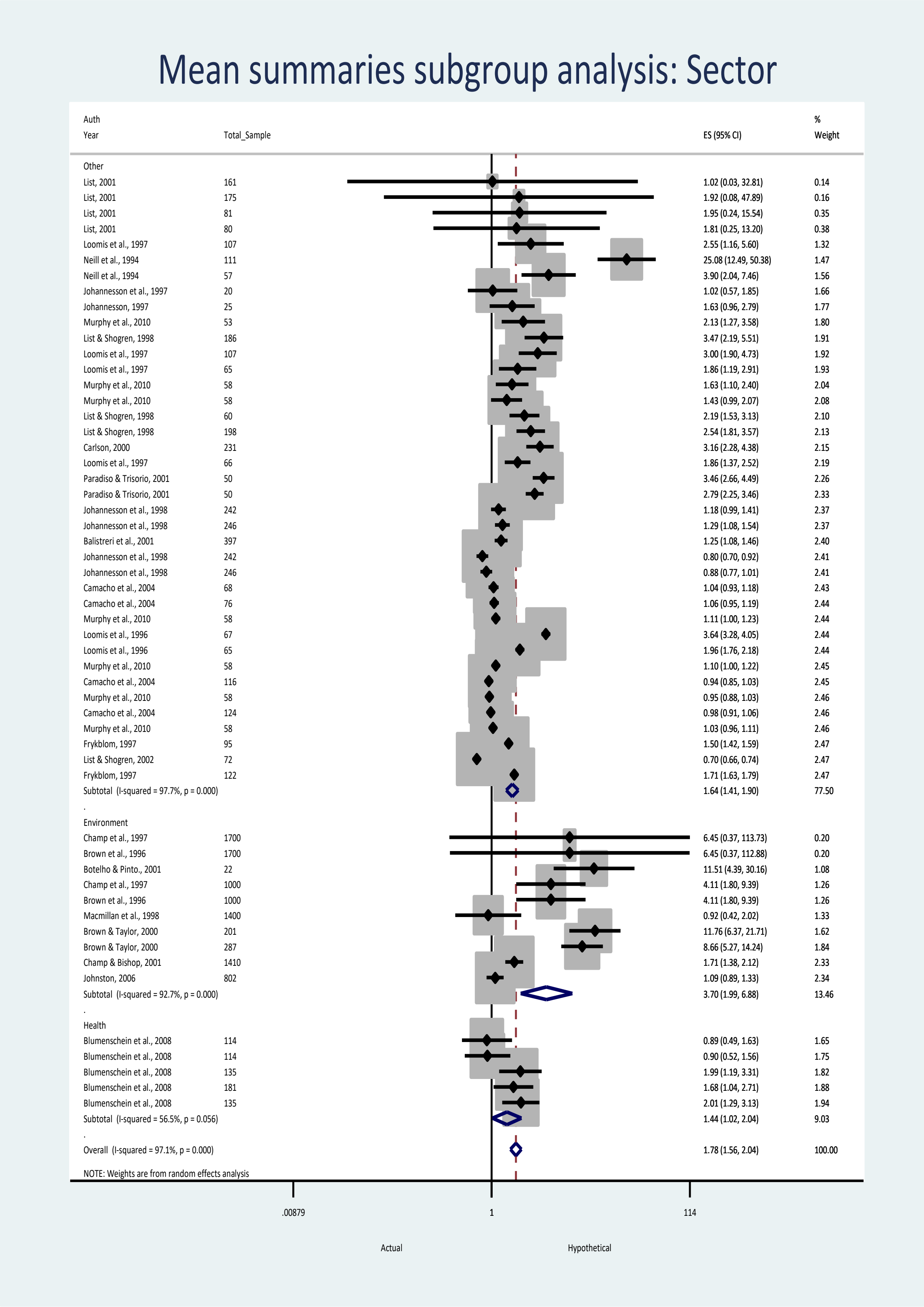


Random effects meta-analysis: Mean summaries

Auth Year	Total_Sample	ES (95% CI)	% Weight
List, 2001	161	1.02 (0.03, 32.81)	0.14
List, 2001	175	1.92 (0.08, 47.89)	0.16
Champ et al., 1997	1700	6.45 (0.37, 113.73)	0.20
Brown et al., 1996	1700	6.45 (0.37, 112.88)	0.20
List, 2001	81	1.95 (0.24, 15.54)	0.35
List, 2001	80	1.81 (0.25, 13.20)	0.38
Botelho & Pinto., 2001	22	11.51 (4.39, 30.16)	1.08
Champ et al., 1997	1000	4.11 (1.80, 9.39)	1.26
Brown et al., 1996	1000	4.11 (1.80, 9.39)	1.26
Loomis et al., 1997	107	2.55 (1.16, 5.60)	1.32
Macmillan et al., 1998	1400	0.92 (0.42, 2.02)	1.33
Neill et al., 1994	111	25.08 (12.49, 50.38)	1.47
Neill et al., 1994	57	3.90 (2.04, 7.46)	1.56
Brown & Taylor, 2000	201	11.76 (6.37, 21.71)	1.62
Blumenschein et al., 2008	114	0.89 (0.49, 1.63)	1.65
Johannesson et al., 1997	20	1.02 (0.57, 1.85)	1.66
Blumenschein et al., 2008	114	0.90 (0.52, 1.56)	1.75
Johannesson, 1997	25	1.63 (0.96, 2.79)	1.77
Murphy et al., 2010	53	2.13 (1.27, 3.58)	1.80
Blumenschein et al., 2008	135	1.99 (1.19, 3.31)	1.82
Brown & Taylor, 2000	287	8.66 (5.27, 14.24)	1.84
Blumenschein et al., 2008	181	1.68 (1.04, 2.71)	1.88
List & Shogren, 1998	186	3.47 (2.19, 5.51)	1.91
Loomis et al., 1997	107	3.00 (1.90, 4.73)	1.92
Loomis et al., 1997	65	1.86 (1.19, 2.91)	1.93
Blumenschein et al., 2008	135	2.01 (1.29, 3.13)	1.94
Murphy et al., 2010	58	1.63 (1.10, 2.40)	2.04
Murphy et al., 2010	58	1.43 (0.99, 2.07)	2.08
List & Shogren, 1998	60	2.19 (1.53, 3.13)	2.10
List & Shogren, 1998	198	2.54 (1.81, 3.57)	2.13
Carlson, 2000	231	3.16 (2.28, 4.38)	2.15
Loomis et al., 1997	66	1.86 (1.37, 2.52)	2.19
Paradiso & Trisorio, 2001	50	3.46 (2.66, 4.49)	2.26
Paradiso & Trisorio, 2001	50	2.79 (2.25, 3.46)	2.33
Champ & Bishop, 2001	1410	1.71 (1.38, 2.12)	2.33
Johnston, 2006	802	1.09 (0.89, 1.33)	2.34
Johannesson et al., 1998	242	1.18 (0.99, 1.41)	2.37
Johannesson et al., 1998	246	1.29 (1.08, 1.54)	2.37
Balistreri et al., 2001	397	1.25 (1.08, 1.46)	2.40
Johannesson et al., 1998	242	0.80 (0.70, 0.92)	2.41
Johannesson et al., 1998	246	0.88 (0.77, 1.01)	2.41
Camacho et al., 2004	68	1.04 (0.93, 1.18)	2.43
Camacho et al., 2004	76	1.06 (0.95, 1.19)	2.44
Murphy et al., 2010	58	1.11 (1.00, 1.23)	2.44
Loomis et al., 1996	67	3.64 (3.28, 4.05)	2.44
Loomis et al., 1996	65	1.96 (1.76, 2.18)	2.44
Murphy et al., 2010	58	1.10 (1.00, 1.22)	2.45
Camacho et al., 2004	116	0.94 (0.85, 1.03)	2.45
Murphy et al., 2010	58	0.95 (0.88, 1.03)	2.46
Camacho et al., 2004	124	0.98 (0.91, 1.06)	2.46
Murphy et al., 2010	58	1.03 (0.96, 1.11)	2.46
Frykblom, 1997	95	1.50 (1.42, 1.59)	2.47
List & Shogren, 2002	72	0.70 (0.66, 0.74)	2.47
Frykblom, 1997	122	1.71 (1.63, 1.79)	2.47
Overall (I-squared = 97.1%, p = 0.000)		1.78 (1.56, 2.04)	100.00

NOTE: Weights are from random effects analysis

Figure 6-2: Forest plot illustrating the subgroup analysis by sector



Mean summaries subgroup analysis: Sector

Auth Year	Total_Sample	ES (95% CI)	% Weight
Other			
List, 2001	161	1.02 (0.03, 32.81)	0.14
List, 2001	175	1.92 (0.08, 47.89)	0.16
List, 2001	81	1.95 (0.24, 15.54)	0.35
List, 2001	80	1.81 (0.25, 13.20)	0.38
Loomis et al., 1997	107	2.55 (1.16, 5.60)	1.32
Neill et al., 1994	111	25.08 (12.49, 50.38)	1.47
Neill et al., 1994	57	3.90 (2.04, 7.46)	1.56
Johannesson et al., 1997	20	1.02 (0.57, 1.85)	1.66
Johannesson, 1997	25	1.63 (0.96, 2.79)	1.77
Murphy et al., 2010	53	2.13 (1.27, 3.58)	1.80
List & Shogren, 1998	186	3.47 (2.19, 5.51)	1.91
Loomis et al., 1997	107	3.00 (1.90, 4.73)	1.92
Loomis et al., 1997	65	1.86 (1.19, 2.91)	1.93
Murphy et al., 2010	58	1.63 (1.10, 2.40)	2.04
Murphy et al., 2010	58	1.43 (0.99, 2.07)	2.08
List & Shogren, 1998	60	2.19 (1.53, 3.13)	2.10
List & Shogren, 1998	198	2.54 (1.81, 3.57)	2.13
Carlson, 2000	231	3.16 (2.28, 4.38)	2.15
Loomis et al., 1997	66	1.86 (1.37, 2.52)	2.19
Paradiso & Trisorio, 2001	50	3.46 (2.66, 4.49)	2.26
Paradiso & Trisorio, 2001	50	2.79 (2.25, 3.46)	2.33
Johannesson et al., 1998	242	1.18 (0.99, 1.41)	2.37
Johannesson et al., 1998	246	1.29 (1.08, 1.54)	2.37
Balistreri et al., 2001	397	1.25 (1.08, 1.46)	2.40
Johannesson et al., 1998	242	0.80 (0.70, 0.92)	2.41
Johannesson et al., 1998	246	0.88 (0.77, 1.01)	2.41
Camacho et al., 2004	68	1.04 (0.93, 1.18)	2.43
Camacho et al., 2004	76	1.06 (0.95, 1.19)	2.44
Murphy et al., 2010	58	1.11 (1.00, 1.23)	2.44
Loomis et al., 1996	67	3.64 (3.28, 4.05)	2.44
Loomis et al., 1996	65	1.96 (1.76, 2.18)	2.44
Murphy et al., 2010	58	1.10 (1.00, 1.22)	2.45
Camacho et al., 2004	116	0.94 (0.85, 1.03)	2.45
Murphy et al., 2010	58	0.95 (0.88, 1.03)	2.46
Camacho et al., 2004	124	0.98 (0.91, 1.06)	2.46
Murphy et al., 2010	58	1.03 (0.96, 1.11)	2.46
Frykblom, 1997	95	1.50 (1.42, 1.59)	2.47
List & Shogren, 2002	72	0.70 (0.66, 0.74)	2.47
Frykblom, 1997	122	1.71 (1.63, 1.79)	2.47
Subtotal (I-squared = 97.7%, p = 0.000)		1.64 (1.41, 1.90)	77.50
.			
Environment			
Champ et al., 1997	1700	6.45 (0.37, 113.73)	0.20
Brown et al., 1996	1700	6.45 (0.37, 112.88)	0.20
Botelho & Pinto., 2001	22	11.51 (4.39, 30.16)	1.08
Champ et al., 1997	1000	4.11 (1.80, 9.39)	1.26
Brown et al., 1996	1000	4.11 (1.80, 9.39)	1.26
Macmillan et al., 1998	1400	0.92 (0.42, 2.02)	1.33
Brown & Taylor, 2000	201	11.76 (6.37, 21.71)	1.62
Brown & Taylor, 2000	287	8.66 (5.27, 14.24)	1.84
Champ & Bishop, 2001	1410	1.71 (1.38, 2.12)	2.33
Johnston, 2006	802	1.09 (0.89, 1.33)	2.34
Subtotal (I-squared = 92.7%, p = 0.000)		3.70 (1.99, 6.88)	13.46
.			
Health			
Blumenschein et al., 2008	114	0.89 (0.49, 1.63)	1.65
Blumenschein et al., 2008	114	0.90 (0.52, 1.56)	1.75
Blumenschein et al., 2008	135	1.99 (1.19, 3.31)	1.82
Blumenschein et al., 2008	181	1.68 (1.04, 2.71)	1.88
Blumenschein et al., 2008	135	2.01 (1.29, 3.13)	1.94
Subtotal (I-squared = 56.5%, p = 0.056)		1.44 (1.02, 2.04)	9.03
.			
Overall (I-squared = 97.1%, p = 0.000)		1.78 (1.56, 2.04)	100.00

NOTE: Weights are from random effects analysis

Figure 6-3: Forest plot illustrating the subgroup analysis by survey setting_Mean summaries

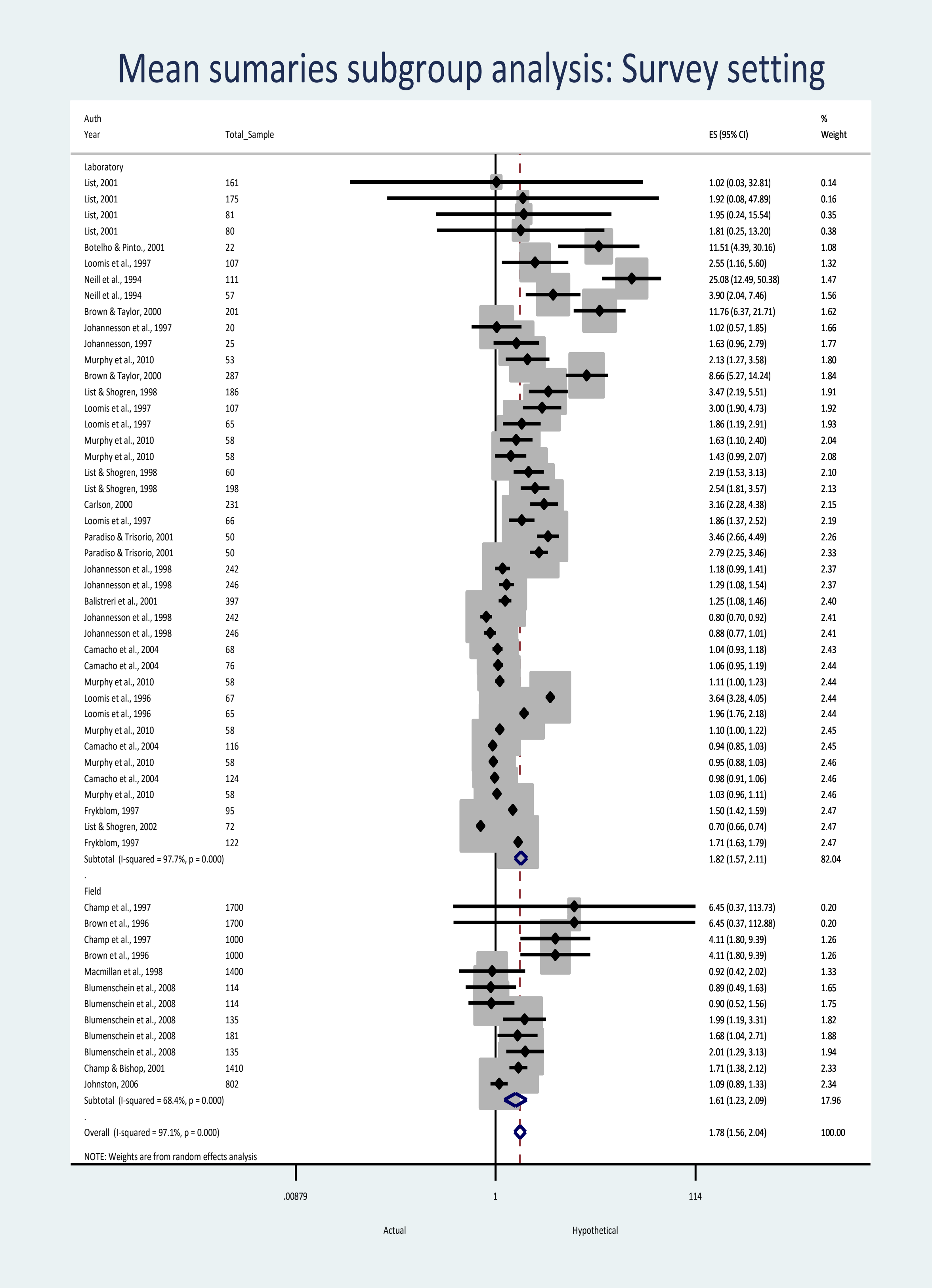
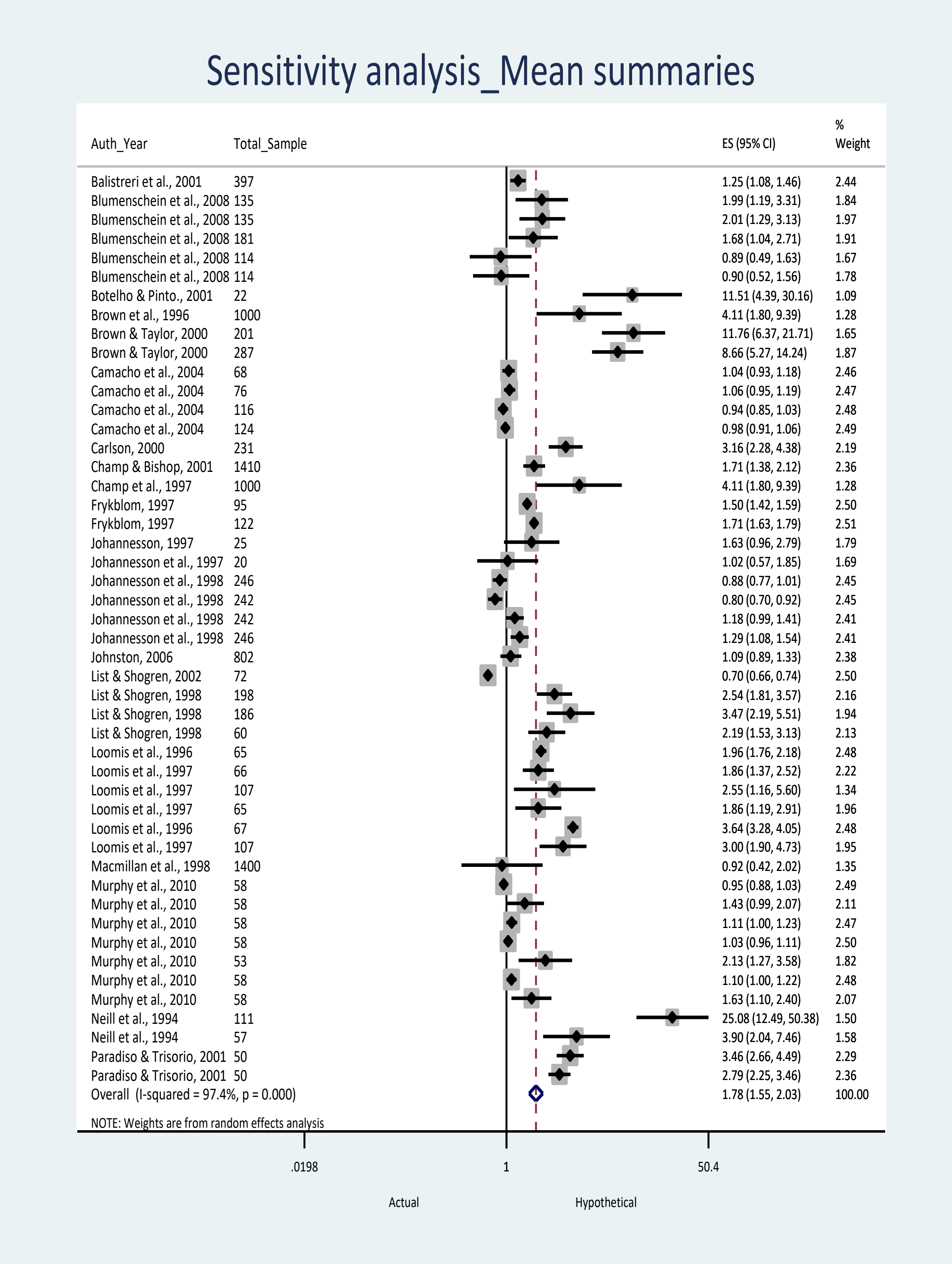


Figure 6-4: Forest plot illustrating results of sensitivity analysis_Mean Summaries



Auth_Year	Total_Sample	ES (95% CI)	% Weight
Balistreri et al., 2001	397	1.25 (1.08, 1.46)	2.44
Blumenschein et al., 2008	135	1.99 (1.19, 3.31)	1.84
Blumenschein et al., 2008	135	2.01 (1.29, 3.13)	1.97
Blumenschein et al., 2008	181	1.68 (1.04, 2.71)	1.91
Blumenschein et al., 2008	114	0.89 (0.49, 1.63)	1.67
Blumenschein et al., 2008	114	0.90 (0.52, 1.56)	1.78
Botelho & Pinto., 2001	22	11.51 (4.39, 30.16)	1.09
Brown et al., 1996	1000	4.11 (1.80, 9.39)	1.28
Brown & Taylor, 2000	201	11.76 (6.37, 21.71)	1.65
Brown & Taylor, 2000	287	8.66 (5.27, 14.24)	1.87
Camacho et al., 2004	68	1.04 (0.93, 1.18)	2.46
Camacho et al., 2004	76	1.06 (0.95, 1.19)	2.47
Camacho et al., 2004	116	0.94 (0.85, 1.03)	2.48
Camacho et al., 2004	124	0.98 (0.91, 1.06)	2.49
Carlson, 2000	231	3.16 (2.28, 4.38)	2.19
Champ & Bishop, 2001	1410	1.71 (1.38, 2.12)	2.36
Champ et al., 1997	1000	4.11 (1.80, 9.39)	1.28
Frykblom, 1997	95	1.50 (1.42, 1.59)	2.50
Frykblom, 1997	122	1.71 (1.63, 1.79)	2.51
Johannesson, 1997	25	1.63 (0.96, 2.79)	1.79
Johannesson et al., 1997	20	1.02 (0.57, 1.85)	1.69
Johannesson et al., 1998	246	0.88 (0.77, 1.01)	2.45
Johannesson et al., 1998	242	0.80 (0.70, 0.92)	2.45
Johannesson et al., 1998	242	1.18 (0.99, 1.41)	2.41
Johannesson et al., 1998	246	1.29 (1.08, 1.54)	2.41
Johnston, 2006	802	1.09 (0.89, 1.33)	2.38
List & Shogren, 2002	72	0.70 (0.66, 0.74)	2.50
List & Shogren, 1998	198	2.54 (1.81, 3.57)	2.16
List & Shogren, 1998	186	3.47 (2.19, 5.51)	1.94
List & Shogren, 1998	60	2.19 (1.53, 3.13)	2.13
Loomis et al., 1996	65	1.96 (1.76, 2.18)	2.48
Loomis et al., 1997	66	1.86 (1.37, 2.52)	2.22
Loomis et al., 1997	107	2.55 (1.16, 5.60)	1.34
Loomis et al., 1997	65	1.86 (1.19, 2.91)	1.96
Loomis et al., 1996	67	3.64 (3.28, 4.05)	2.48
Loomis et al., 1997	107	3.00 (1.90, 4.73)	1.95
Macmillan et al., 1998	1400	0.92 (0.42, 2.02)	1.35
Murphy et al., 2010	58	0.95 (0.88, 1.03)	2.49
Murphy et al., 2010	58	1.43 (0.99, 2.07)	2.11
Murphy et al., 2010	58	1.11 (1.00, 1.23)	2.47
Murphy et al., 2010	58	1.03 (0.96, 1.11)	2.50
Murphy et al., 2010	53	2.13 (1.27, 3.58)	1.82
Murphy et al., 2010	58	1.10 (1.00, 1.22)	2.48
Murphy et al., 2010	58	1.63 (1.10, 2.40)	2.07
Neill et al., 1994	111	25.08 (12.49, 50.38)	1.50
Neill et al., 1994	57	3.90 (2.04, 7.46)	1.58
Paradiso & Trisorio, 2001	50	3.46 (2.66, 4.49)	2.29
Paradiso & Trisorio, 2001	50	2.79 (2.25, 3.46)	2.36
Overall (I-squared = 97.4%, p = 0.000)		1.78 (1.55, 2.03)	100.00

6.3.1.2: Comparisons presenting summaries as percentages

The log odds ratio of the actual and hypothetical percentages was used in the random effects meta-analysis. The forest plot shows that respondents were more likely to say “yes” in the hypothetical survey than they were in the actual survey (figure 6.2). The pooled odds ratio from the studies presenting percent summaries was 2.37 (range 1.93 – 2.80). This means that the odds of saying “yes” in the hypothetical survey were more than double the odds of saying “yes” in the actual survey. Stated differently, the odds of saying yes in the actual survey were 137% lower than they were in the hypothetical survey. In two comparisons from one study (Vossler et al., 2003) the odds of saying yes in the hypothetical survey were less than the odds of saying yes in the actual survey. Overall, the level of heterogeneity (I^2) was high at 90.2% and statistically significant (p=0.000). This means that the variation in the pooled studies could not be attributed to chance alone. This heterogeneity is explored in the meta-regressions presented in a later section.

In the subgroup analysis, differences in the level of heterogeneity were noted in the sector and survey setting. The level of heterogeneity was very high and significant for studies from the environment sector (93.25%, p=0.000). Heterogeneity in the other and health sectors was considerably low and insignificant (35.2%, p=0.117 & 21.9%, p=0.211 respectively). This suggests that the variation in the effect sizes from these studies could be attributed to chance alone (Figure 6.6). The level of heterogeneity was also lower in studies conducted in a laboratory setting (I^2: 40.9% but this was not significant (p=0.133) compared to field studies (I^2=91.1%, p=0.000). These differences by survey setting could be attributed to the few laboratory studies (Figure 6.7). The differences in the levels of heterogeneity were not significantly for the other study attributes.

Sensitivity Analyses

For the 4 studies which reported zero log odds ratios, the zero values in the actual survey were replaced with of 0.001 and 0.0001. In both cases, the results were the same with an overall effect size is 2.36 (range 1.96-2.85). The variation in effect size which is attributable to heterogeneity is 90.5% and this is statistically significant (p=0.001) (Figure 6.8). These results are not very different from the results obtained when only studies with non-zero log odds ratios are analyzed.

Figure 6-5: Forest plot of log odd-ratios for comparisons reporting percentage summaries

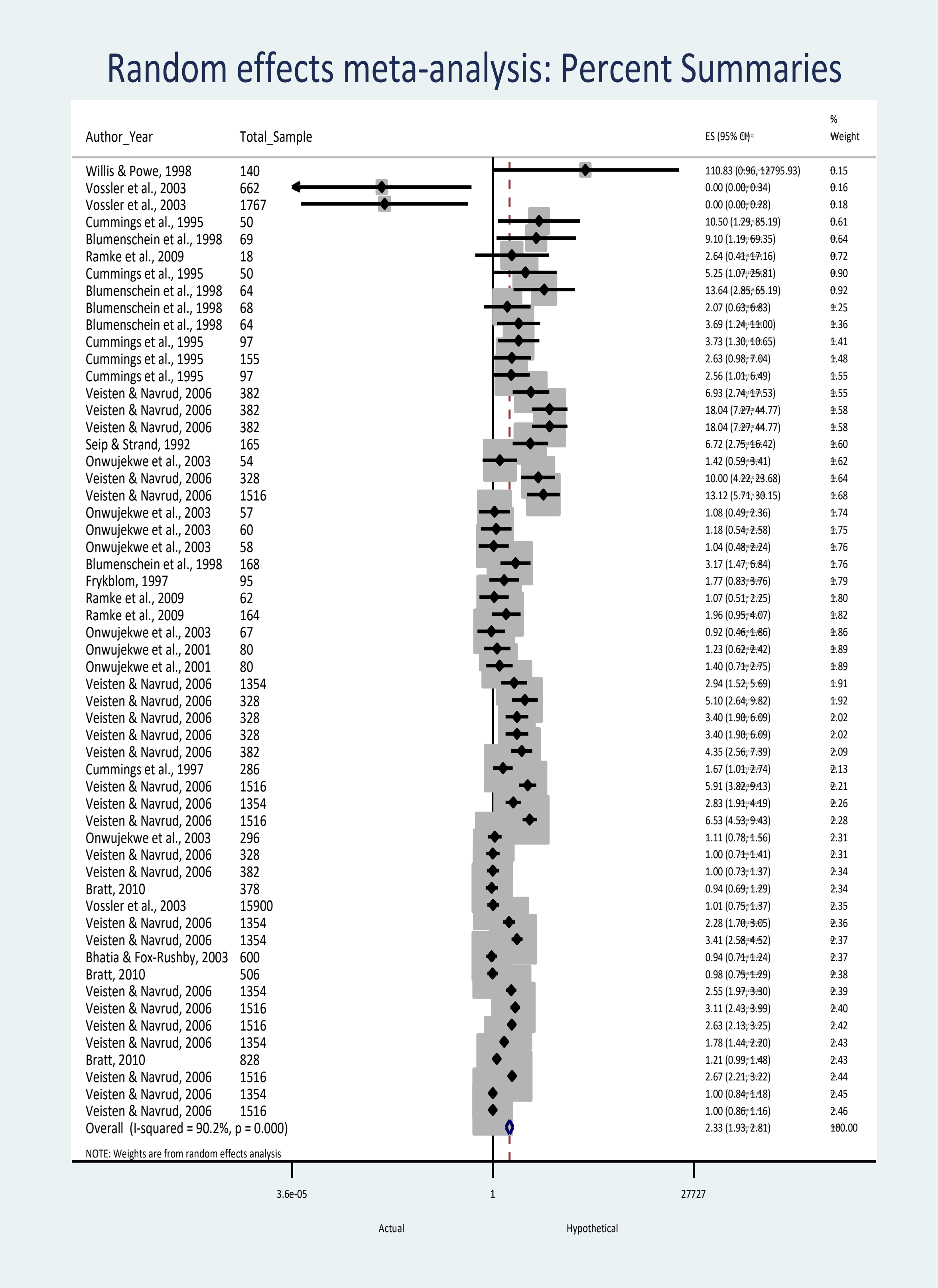


Figure 6-6: Forest plot illustrating the subgroup analysis by sector _ Percentage summaries

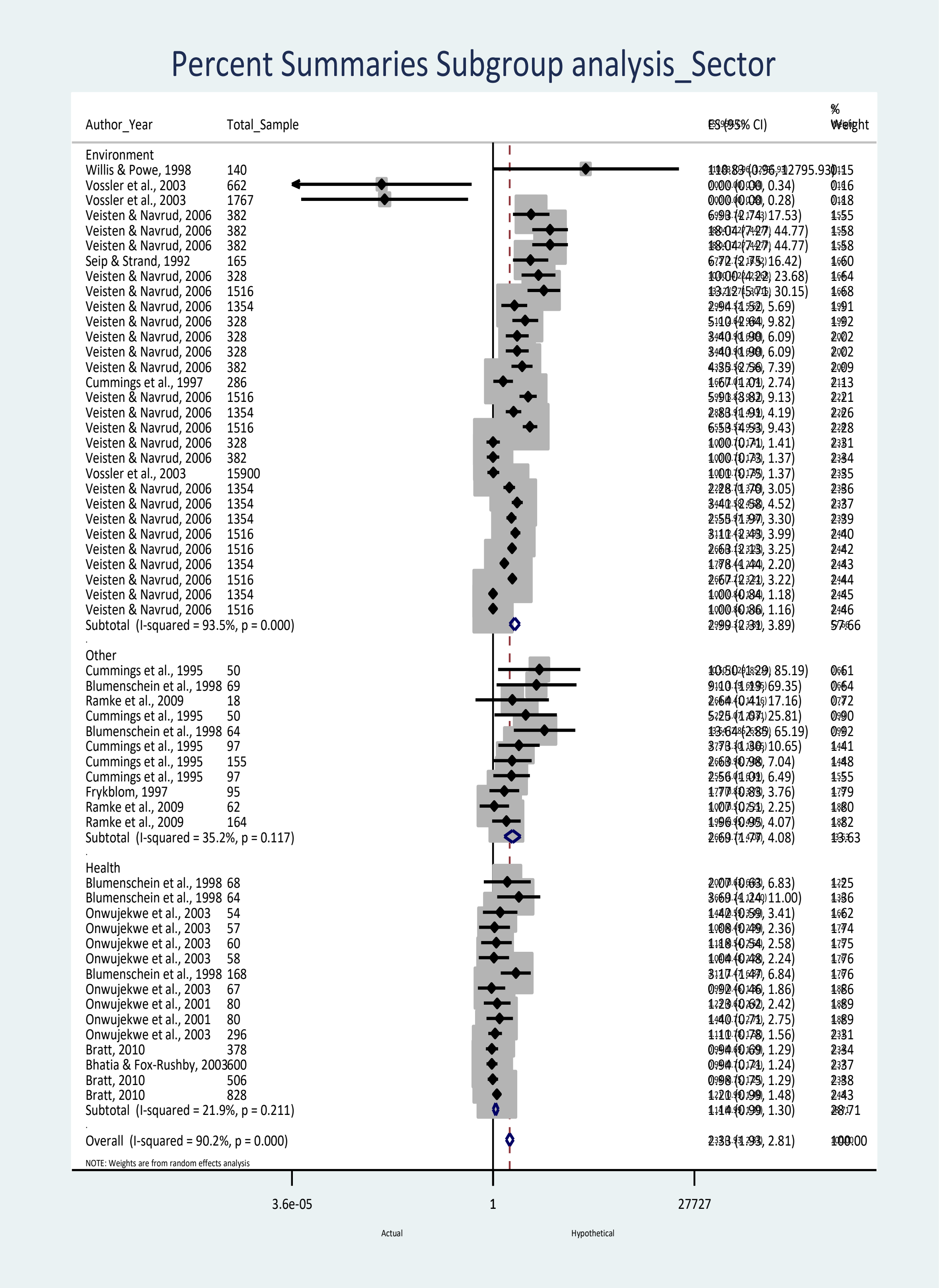


Figure 6-7: Forest plot illustrating the subgroup analysis by survey setting _ Percentage summaries

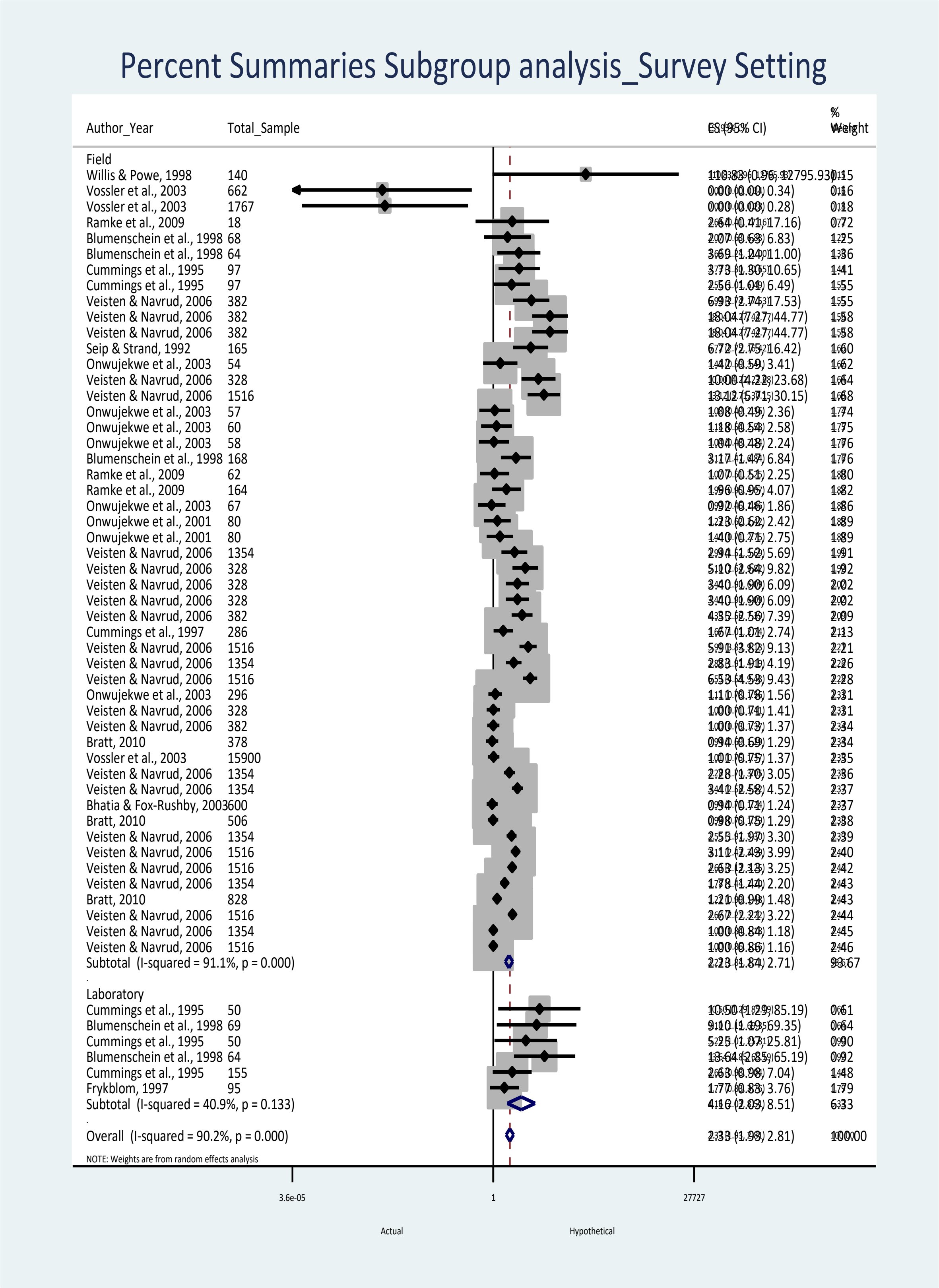
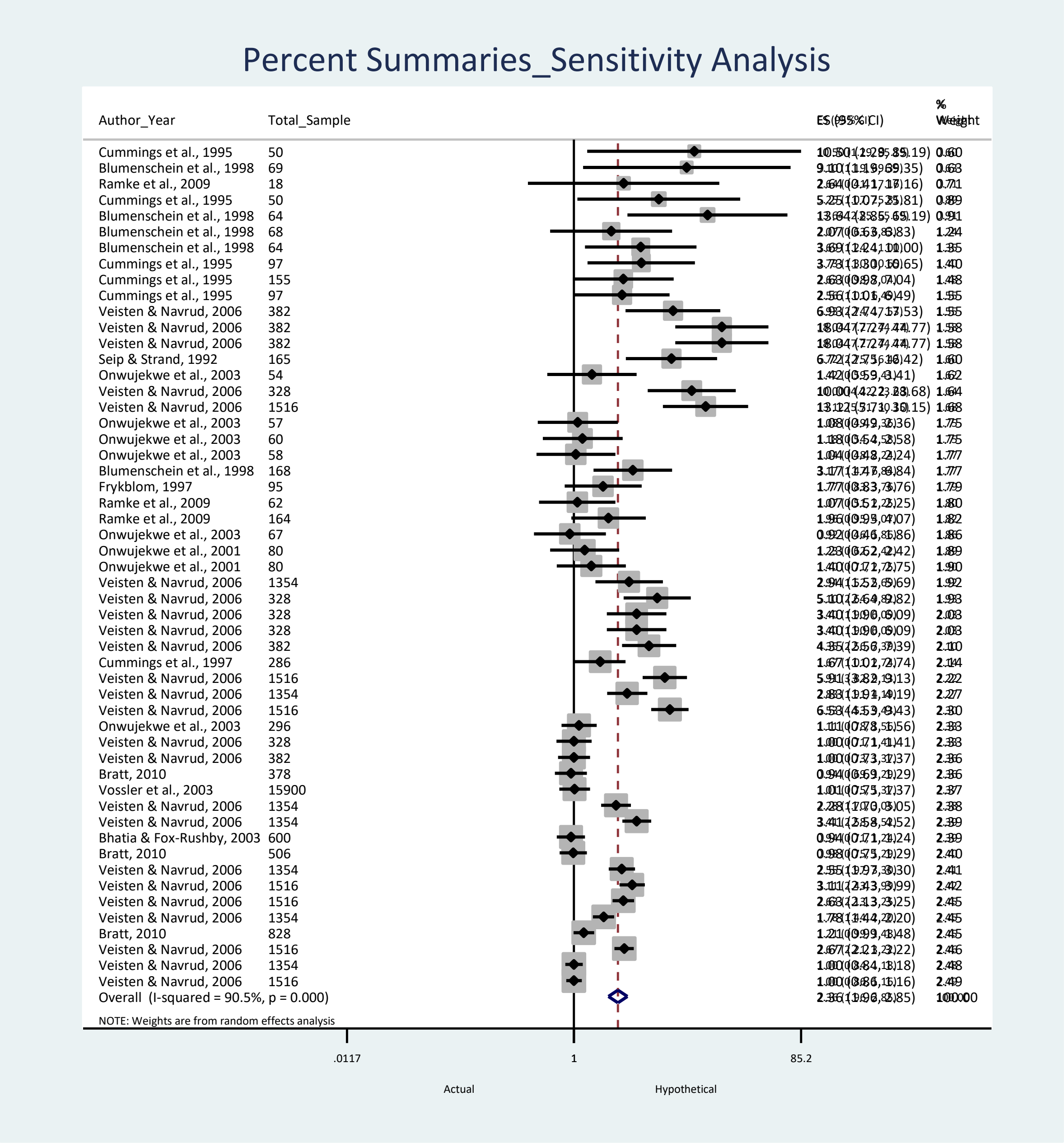


Figure 6-8: Forest plot illustrating results of sensitivity analysis _Percentage Summaries



Percent Summaries_Sensitivity Analysis

Author_Year	Total_Sample	ES (95% CI)	% Weight
Cummings et al., 1995	50	10.50 (1.29, 85.19)	0.60
Blumenschein et al., 1998	69	9.10 (1.19, 69.35)	0.63
Ramke et al., 2009	18	2.64 (0.41, 17.16)	0.71
Cummings et al., 1995	50	5.25 (1.07, 25.81)	0.89
Blumenschein et al., 1998	64	13.64 (2.85, 65.19)	0.91
Blumenschein et al., 1998	68	2.07 (0.63, 6.83)	1.24
Blumenschein et al., 1998	64	3.69 (1.24, 11.00)	1.35
Cummings et al., 1995	97	3.73 (1.30, 10.65)	1.40
Cummings et al., 1995	155	2.63 (0.98, 7.04)	1.48
Cummings et al., 1995	97	2.56 (1.01, 6.49)	1.55
Veisten & Navrud, 2006	382	6.93 (2.74, 17.53)	1.55
Veisten & Navrud, 2006	382	18.04 (7.27, 44.77)	1.58
Veisten & Navrud, 2006	382	18.04 (7.27, 44.77)	1.58
Seip & Strand, 1992	165	6.72 (2.75, 16.42)	1.60
Onwujekwe et al., 2003	54	1.42 (0.59, 3.41)	1.62
Veisten & Navrud, 2006	328	10.00 (4.22, 23.68)	1.64
Veisten & Navrud, 2006	1516	13.12 (5.71, 30.15)	1.68
Onwujekwe et al., 2003	57	1.08 (0.49, 2.36)	1.75
Onwujekwe et al., 2003	60	1.18 (0.54, 2.58)	1.75
Onwujekwe et al., 2003	58	1.04 (0.48, 2.24)	1.77
Blumenschein et al., 1998	168	3.17 (1.47, 6.84)	1.77
Frykblom, 1997	95	1.77 (0.83, 3.76)	1.79
Ramke et al., 2009	62	1.07 (0.51, 2.25)	1.80
Ramke et al., 2009	164	1.96 (0.95, 4.07)	1.82
Onwujekwe et al., 2003	67	0.92 (0.46, 1.86)	1.86
Onwujekwe et al., 2001	80	1.23 (0.62, 2.42)	1.89
Onwujekwe et al., 2001	80	1.40 (0.71, 2.75)	1.90
Veisten & Navrud, 2006	1354	2.94 (1.52, 5.69)	1.92
Veisten & Navrud, 2006	328	5.10 (2.64, 9.82)	1.93
Veisten & Navrud, 2006	328	3.40 (1.90, 6.09)	2.03
Veisten & Navrud, 2006	328	3.40 (1.90, 6.09)	2.03
Veisten & Navrud, 2006	382	4.35 (2.56, 7.39)	2.10
Cummings et al., 1997	286	1.67 (1.01, 2.74)	2.14
Veisten & Navrud, 2006	1516	5.91 (3.82, 9.13)	2.22
Veisten & Navrud, 2006	1354	2.83 (1.91, 4.19)	2.27
Veisten & Navrud, 2006	1516	6.53 (4.53, 9.43)	2.30
Onwujekwe et al., 2003	296	1.11 (0.78, 1.56)	2.33
Veisten & Navrud, 2006	328	1.00 (0.71, 1.41)	2.33
Veisten & Navrud, 2006	382	1.00 (0.73, 1.37)	2.36
Bratt, 2010	378	0.94 (0.69, 1.29)	2.36
Vossler et al., 2003	15900	1.01 (0.75, 1.37)	2.37
Veisten & Navrud, 2006	1354	2.28 (1.70, 3.05)	2.38
Veisten & Navrud, 2006	1354	3.41 (2.58, 4.52)	2.39
Bhatia & Fox-Rushby, 2003	600	0.94 (0.71, 1.24)	2.39
Bratt, 2010	506	0.98 (0.75, 1.29)	2.40
Veisten & Navrud, 2006	1354	2.55 (1.97, 3.30)	2.41
Veisten & Navrud, 2006	1516	3.11 (2.43, 3.99)	2.42
Veisten & Navrud, 2006	1516	2.63 (2.13, 3.25)	2.45
Veisten & Navrud, 2006	1354	1.78 (1.44, 2.20)	2.45
Bratt, 2010	828	1.21 (0.99, 1.48)	2.45
Veisten & Navrud, 2006	1516	2.67 (2.21, 3.22)	2.46
Veisten & Navrud, 2006	1354	1.00 (0.84, 1.18)	2.48
Veisten & Navrud, 2006	1516	1.00 (0.86, 1.16)	2.49
Overall (I-squared = 90.5%, p = 0.000)		2.36 (1.96, 2.85)	100.00

NOTE: Weights are from random effects analysis

6.3.2 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. Of the percentage summaries, four reported a zero value in the actual survey results, generating an odds ratio of zero. The descriptive statistics for all the variables included in the regression models is presented in (Appendix 16). Base and reduced models are run separately for the mean and percentage summaries. In both models, independent predictors are entered into the model as binary predictors. The models are also run with the sector, class and purpose of good variables entered as dummy variables representing the distinct categories. Both univariate and meta-regressions are considered, and the results presented separately.

For all the presented models, the linktest estimate was not significant, indicating that the models were correctly fitted. In interpreting 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. The univariate regression results for the mean and percent summaries are discussed below with the meta-regression results presented after. The model outputs are combined in table 6.1 for mean summaries and table 6.2 for percentage summaries. For each of these, the results of the regression analysis using the dummy variables for sector, class and purpose of good are also discussed. The base and reduced model outputs where dummy variables are included in the meta-regressions are combined for mean summaries (Appendix 17) and percentage summaries (Appendix 18). In the last section, the univariate and meta-regression results are compared and discussed.

6.3.2.1: Univariate meta-regression

Univariate meta-regressions with binary independent variables

In the first analysis, the independent variables were entered into the regression models as binary variables with the log ratio as the dependent variable. The results are discussed below separately for mean and percentage summaries.

Mean summaries

Among the general study variables, goods in the environment sector are positively and significantly related to the log ratio (1.056, $p<0.001$) while those in the health and other sectors are negative but significantly related to the log ratio (health: 0.386, p=0.018; other sector: 0.711, p<0.001) (table 6.1). Pure public goods are positive and significantly related to the log ratio (0.668, p=0.002) while those in the private and other sectors are negatively related to the log ratio (pure private: 0.310, p=0.074; quasi-private goods: 0.285, p=0.173). All the categories in the purpose of good are significantly related to the log ratio of hypothetical and actual WTP values. While conservation goods are positively related (1.022, p<0.001) to the log ratio, goods in the other purposes are negatively related (prevention: 0.409, p=0.006; other purposes: 0.7333, p<0.001).

The duration between the two surveys and the type of comparisons are negatively related to the log ratio but these are not significant. The log ratios are lower when the surveys are held concurrently (0.0393, p=0.852), with a between-sample comparison (0.125, p=0.491) and when respondents are given money for participation in the surveys (0.292, p=0.105). The log ratio is also lower with cash payments and this is significant (0.543, p=0.015). The log ratio is positively and significantly related to the payment duration and survey setting. It is higher when a one-off payment is elicited compared to other regular payments (0.479, p=0.011) and when surveys are held in the field as opposed to laboratories (0.316, p=0.091).

Among the variables comparing differences between the hypothetical and actual surveys at a composite level, the log ratio is significantly lower when student samples and users or potential users of the valuation good are used in both hypothetical and actual surveys of WTP (0.111, p=0.541 and 0.472, p=0.140) respectively but these are not statistically significant. The ratio is also lower when the same sample selection (0.572, p=0.007) and mode of administration (0.734, p=0.059) methods are used in the two surveys but higher (0.0558, p=0.775) with the same elicitation method. When the different sample selection categories are investigated, the log ratios are still lower when the same methods are used in the two surveys (random sampling: 0.601, p<0.001; purposive sampling: 0.001, p=0.994; convenience sampling: 0.189, p=0.278). The different administration modes are also significant, with mail surveys in both surveys generating a higher log ratio (0.385, p=0.071) while in-person surveys generating a lower log ratio (0.577, p=0.004).

Under elicitation methods, the use of bidding and payment card methods leads to significantly lower log ratios (0.310, p=0.001 and 0.719, p<0.001 respectively). When auctions and dichotomous choice methods are used in both surveys the log ratios are also lower (0.067, p=0.7112 and 0.097, p=0.569). However, the use of open ended methods to elicit WTP values in both hypothetical and actual surveys leads to significantly higher log ratios (0.905, p<0.001).

Percentage summaries

The log odds ratio was used as the dependent variable in this analysis. The log ratios of hypothetical and actual surveys are higher when studies are conducted in high income countries (0.835, p=0.006). Among the sector variables, the log ratios for goods in the health sector are positive and significantly lower (0.701, p=0.024) while those in the other sector are positive and significantly higher (0.615, p=0.084). Goods in the environment sector are higher but this is not statistically significant (0.162, p=0.690). Among the class of goods, pure private and quasi-private goods have lower log ratios, and these are not statistically significant (0.162, p=0.690 and 2.475, p=0.271). However, pure public goods have positive and statistically significant log ratios (0.822, p=0.040).

The log ratios for goods in the other and health sectors are statistically significant with lower ratios for prevention goods (0.694, p=0.038) and higher ratios for goods in other sectors (1.331, p=0.006). Conservation goods have lower ratios (0.095, p=0.824) but this difference is not statistically significant. The log ratios are positive and statistically significant when hypothetical and actual surveys are conducted concurrently (0.930, p=0.020), one-off payments elicited (3.364, p=0.054) and when respondents are given money for the purchase of the valuation commodity in the actual survey (0.908, p=0.011). The ratios are also higher when cash payment vehicles are used (0.174, p=0.691) and when money is given to respondents for participation in the surveys (0.333, p=0.395). However, log ratios are significantly lower when between sample comparisons are used (1.411, p=0.083) and when surveys are held in a field setting (1.029, p=0.010).

In comparing attributes which are similar in the hypothetical and actual surveys, log ratios are positive and statistically significant with the use of student samples (1.029, p=0.010), users or potential users (0.864, p<0.001), sample selection methods

(0.737, p=0.004) and administration methods (3.364, p=0.054) in both surveys. The log ratios are lower when the same elicitation methods are used in both the hypothetical and actual surveys of WTP (0.572, p=0.611). The use of the different sample selection methods in both the hypothetical and actuals surveys is significant. The ratios are positive and significant when purposive (0.976, p=0.027) and convenience (1.029, p=0.010) samples are used in both hypothetical and actual surveys and lower when the random sampling methods are used (1.482, p=0.015). The log ratios are positive when mail (0.813, p=0.035) and in-person surveys (0.095, p=0.824) are used in both surveys and this difference is statistically significant with mail surveys. The log ratios are higher, and this is statistically significant when open ended methods are used (0.936, p=0.028). However, the log ratios are lower with the bidding (0.782, p=0.002) and dichotomous choice (0.447, p=0.289) methods in both hypothetical and actual surveys of WTP.

Univariate meta-regression with sector, class and purpose of good as dummy variables

Univariate meta-regressions were run with sector, class and purpose variables entered as dummies to represent the distinct categories. These results are included in the regression output in table 6.1 for mean summaries and table 6.2 for percent summaries and discussed separately below. The results presented in the base and reduced models in table 6.1 do not include the dummy variable analysis for these variables.

Mean summaries

With health as the reference category, the log ratios for goods in the environment sector were positive and significantly different from the health sector goods (table 6.1)). In this, the log ratio for a good in the environment sector is on average 1.05 points higher than the log ratio for a good in the health sector (p<0.001). Goods in the other sector generally had lower log ratios when compared to the health sector (0.006, p= 0.979). Pure public goods have log ratios which are higher than pure private goods and this is statistically significant (0.64, p=0.001). However, quasi-private goods have lower log ratios when compared to quasi-private goods (0.06, p=0.760) but this difference is not statistically significant. With prevention purpose as the reference category, conservation goods have a higher and significant log ratio

(1.02, p<0.001) while goods which are classified in the other (non-prevention and non-conservation goods) have lower log ratios though this is not significant (0.004. p=0.985).

Percentage summaries

With health as the reference category, the log ratios for goods in the environment and other sectors were both positive but not significant (Environment: 0.58, p=0.240; Other: 1.00, p=0.111) (table 6.2). Pure public goods have log ratios which are higher than pure private goods, but this is not statistically significant (0.528, p=0.400). However, quasi-private goods have lower log ratios when compared to quasi-private goods and this difference is significant (2.211, p=0.006). With prevention purpose as the reference category, conservation goods have a higher log ratio (0.411, p=0.366) while goods which are classified in the other (non-prevention and non-conservation goods) have higher and significantly different log ratios (1.575, p=0.013).

6.3.2.2: Multiple meta-regression

Step-wise backward regression models were run with the independent variables entered into the model as binary variables. These results are presented separately for mean (table 6.1) and percentage (table 6.2) summaries and discussed in the next section. The model outputs for the regressions with dummy variables representing the distinct categories are provided in Appendices 17 and 18 for mean and percentage summaries respectively and discussed in the section that follows.

Multiple Meta-regressions with binary independent variables

Mean summaries

For both the base and reduced models the regressions weighted by the study fit the data for r^2 of 0.68 and 0.65 respectively. In both models, the log ratios are lower and significant with health sector goods, those that are used for prevention or other purposes. The log ratios are also lower and significant when random samples, mail and in-person interviews are used in both hypothetical and actual surveys of WTP. However, the log ratios are higher in both the base and reduced models when potential users of the valuation good, auction and open-ended methods are used in both the hypothetical and actual surveys of WTP. The log ratios are also high with the use of the dichotomous choice methods in both hypothetical and actual surveys, but this difference is only significant in the reduced model. The signs are reversed in the base and reduced models for the valuation of private goods which is negative in the base model but positive and significant in the reduced model.

Percentage summaries

The regressions using the log odds ratio as the dependent variable and weighted by the study fit the data very well. The R^2 from the base and reduced models are 0.88 and 0.86 respectively. Negative and significant coefficients are observed in both models for quasi-private goods, the use of cash fee payment vehicle and the use of bidding, dichotomous choice and open ended WTP elicitation methods in both hypothetical and actual surveys of WTP. Although the coefficient was not significant in the base model, the log ratios are lower when the hypothetical and actual surveys of WTP are conducted concurrently. In both models, the log ratios are higher when the study is conducted in high income countries, in the environment sector, when student samples are used in both hypothetical and actual surveys and when the selection of the samples in both surveys is conducted using random and purposive sampling methods. Other significant variables in the base model include valuation of conservation goods and the use of mail surveys in both hypothetical and actual surveys of WTP which are both associated with lower log ratios. The reduced model also identifies the use of in-person surveys as being positive and significant while eliciting one-off payments are associated with lower log ratios.

Multiple meta-regressions with sector, class and purpose of good as dummy variables

Mean summaries

In the base model, none of the sector variables is significant. When compared to the health sector, the environment sector goods generate higher log ratios (0.0001, p=1.000) while goods from the other sector have lower log ratios (0.45, p=0.166). However, with the environment sector as the reference category in the reduced model, both the health and other sectors generate lower log ratios, and these are significant (health 1.29, p<0.001; Other 1.37, p<0.001). With pure private goods as the reference category, pure-public goods have lower log ratios (0.19, p=0.510) while quasi private goods have higher log ratios (0.02, p=0.919). However, none of these are significant and the reduced model does not pick any of them. In the base model, when compared to prevention goods, only conservation goods have a positive and significant sign. The log ratio of conservation goods is higher than prevention goods by a factor of 1.03 (p=0402). None of the categories in this variable are picked in the reduced model.

Variables which have positive and significant variables in both the base and reduced models include having respondents who are users or potential users in both surveys, the use of auction and open-ended methods and having one-off elicitation methods. On the other side, the use of random samples, mail administration methods and in-person surveys in both surveys leads to lower and significant log ratios in both the base ad reduced models. The use of dichotomous choice and open-ended methods in both surveys leads to higher log ratios but these are only significant in the reduced model while the bidding method generates significantly lower log ratios in the reduced model but positive log ratios in the base model.

The coefficient signs are reversed in the base and reduced models for valuations in the environment sector and the use of bidding methods in both the hypothetical and actual surveys of WTP. In both cases, the log ratios are higher in the base models but significantly lower in the reduced models.

Percent summaries

In the base model, with the health sector as the reference category, the log ratio for goods in the environment sector are lower than those in the health sector but this is not statistically significant (1.142, p=0.168). The log ratios for goods in the health sector are higher than those in the health sector but this too is not significant (0.401, p=0.141). None of these categories are picked in the reduced model. With pure public goods as the reference category, the log ratios for pure private goods are higher (10.96) and this is statistically significant (p<0.001). These categories are not picked in the reduced model. With conservation as the reference category, prevention goods have log ratios which are lower and statistically significant (9.99, p<0.001=). Goods in the other sector have comparatively lower log rations too (0.731) but this is not statistically significant (p=0.280). None of these categories are picked in the reduced model.

The coefficient for the income level of the country in which the study is conducted is positive and significant in both the base and reduced models. For both models, the log ratios are negative when between sample comparisons are used, and this is significant in the reduced model. The coefficient signs are reversed when random samples are used in both hypothetical and actual surveys of WTP with the log ratio positive in the base but negative in the reduced models and both are significant. In the reduced model, the use of in-person methods in both surveys leads to positive and significant log ratios. In the base model, positive and significant log ratios are associated with valuation of pure private goods and the use of purposive sampling methods in both the hypothetical and actual surveys. In the same model, negative and significant log ratios are observed when goods intended for prevention purposes are used. Lower and significant coefficients are also associated with the use of bidding, dichotomous and open-ended methods in both the hypothetical and actual surveys of WTP.

Comparison of variables across univariate and multivariate regression models

Mean summaries

Elicitation of one-off payments and the use of open ended surveys in both hypothetical and actual surveys of WTP are associated with positive and significant log ratios which are stable across the three models. Variables with coefficients that are associated with lower and significant log ratios across the three models are; valuation of health sector and goods meant for prevention and other purposes, the use of random sampling selection methods and in-person surveys in both surveys.

The coefficient and signs for the one-off payment methods, open ended elicitation methods, in-person surveys and random sampling selection methods are also the same when dummy variables are used in the models.

Percent summaries

Across the univariate, base and reduced models, the income level of the country where a study is conducted and the use of student samples in both the hypothetical and actual surveys of WTP are positive and significant. The income level remains positive and significant in the model with dummy variables for class, sector and purpose of good. The use of bidding and dichotomous choice methods is associated with negative and significant coefficients. None of the significant variables identified in this summary are similar across the mean and percentage summaries.

Table 6-1: Meta-regression results for comparisons presenting mean summaries

Regressions	**Univariate meta-regression**	**Multiple meta-regressions**	
Variables	**Coefficient (SE)**	**Base model** **Coefficient (SE)**	**Reduced model** **Coefficient (SE)**
General study attributes			
High Income Country	-	-	-
Sector [entered as dummy variables: reference category: Health]			
Environment Sector	1.051***(0.254)		
Other Sector	-0.00611 (0.233)		
Sector entered as binary variables			
Environment Sector	1.056*** (0.179)	0.442 (0.527)	
Other Sector	-0.711*** (0.172)	-	-0.807*** (0.197)
Health Sector	-0.386** (0.160)	-0.318** (0.156)	-0.898*** (0.199)
Class of good [entered as dummy variables: reference category: Pure private good]			
Pure Public Good	0.647*** (0.196)		
Quasi-Private Good	-0.0622 (0.203)		
Class of good entered as binary variables			
Pure Public Good	0.668*** (0.206)	-0.229 (0.289)	-
Pure Private Good	-0.310* (0.171)	-0.046 (0.256)	0.610*** (0.213)
Quasi-Private Good	-0.285 (0.207)	-	0.614*** (0.186)
Purpose of good [entered as dummy variables: Reference category: Prevention]			
Conservation Purpose	1.026***(0.258)		
Other Purposes (Besides, Prevention and Conservation)	0.00460 (0.237)		
Purpose of good entered as binary variables			
Conservation Purpose	1.022*** (0.186)	-0.464 (0.575)	
Other Purposes (Besides, Prevention and Conservation)	-0.733*** (0.174)	-1.507*** (0.352)	-0.663*** (0.162)
Prevention Purpose	-0.409*** (0.146)	-0.787*** (0.173)	-0.632*** (0.187)
Duration between surveys			
Hypothetical and Actual surveys held concurrently	-0.0393 (0.210)	0.293 (0.234)	
Payment duration			
One-off payment elicited	0.479** (0.185)	1.165*** (0.395)	1.460*** (0.359)

Regressions	**Univariate meta-regression**	**Multiple meta-regressions**	
Variables	**Coefficient (SE)**	**Base model Coefficient (SE)**	**Reduced model Coefficient (SE)**
Payment vehicle			
Cash fee payment vehicle	-0.543** (0.218)	-0.235 (0.303)	
Type of comparison			
Between sample comparisons	-0.125 (0.180)	-0.175 (0.207)	
Survey setting			
Surveys held in a field setting	0.316* (0.185)	-0.323 (0.343)	
Money effects			
Money given for participation in either survey	-0.292 (0.178)	-0.0914 (0.192)	
Comparisons between study attributes in hypothetical and actual surveys			
Sample type [Student or not]			
Student sample in both surveys	-0.111 (0.181)	0.430 (0.381)	
Sample type [Potential user or not]			
Respondent a potential user in both surveys	-0.472 (0.317)	0.494* (0.269)	0.504***(0.163)
Sample selection			
Same sample selection method in both surveys	-0.572*** (0.205)		
Random sampling in both	-0.601*** (0.160)	-1.579*** (0.473)	-1.220*** (0.312)
Purposive sampling in both	-0.00147 (0.180)	-0.265 (0.280)	
Convenience sampling in both	-0.189 (0.173)	-0.422 (0.351)	
Administration mode			
Same mode of administration in both surveys	-0.734* (0.384)	-	
Mail administration in both surveys	0.385* (0.211)	-0.865** (0.360)	-1.029*** (0.179)
In-person surveys in both surveys	-0.577*** (0.195)	-0.802*** (0.283)	-0.631*** (0.181)
Elicitation method			
Same elicitation method in both surveys	0.0558 (0.195)	-	
Auction method in both surveys	-0.067 (0.182)	0.886** (0.439)	0.793*** (0.183)
Bidding method in both surveys	-0.310*** (0.0882)	0.485 (0.756)	
Dichotomous choice methods in both surveys	-0.097 (0.171)	0.430 (0.276)	0.397*** (0.146)
Open ended methods in both surveys	0.905*** (0.234)	0.656** (0.301)	0.631*** (0.129)
Payment card method in both surveys	-0.719*** (0.136)	-0.0347 (0.262)	

Regressions	Univariate meta-regression	Multiple meta-regressions	
Variables	Coefficient (SE)	Base model Coefficient (SE)	Reduced model Coefficient (SE)
Observations	84	84	84
R-squared		0.6836	0.659

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

Table 6-2: Meta-regression results for comparisons presenting percent summaries

Variables	Univariate meta-regressions	Multiple meta-regressions	
	Coefficient (SE)	Base Coef. (SE)	Reduced Coef. (SE)
General study attributes			
High Income Country	0.835*** (0.291)	1.919** (0.749)	1.349***(0.476)
Sector [entered as dummy variables: reference category: Health]			
Environment Sector	0.588 (0.495)		
Other Sector	1.007 (0.622)		
Sector entered as binary variables			
Environment Sector	0.162 (0.405)	9.412*** (0.574)	9.577***(1.219)
Other Sector	0.615* (0.350)	-	
Health Sector	-0.701** (0.303)	-0.401 (0.268)	
Class of good [entered as dummy variables: reference category: Pure private good]			
Pure Public Good	0.528 (0.400)		
Quasi-Private Good	-2.211*** (0.775)		
Class of good entered as binary variables			
Pure Public Good	0.822** (0.391)	-	
Pure Private Good	-0.162 (0.405)	-	
Quasi-Private Good	-2.475 (2.223)	-10.96*** (1.088)	-10.32***(1.427)
Purpose of good [entered as dummy variables: Reference category: Prevention]			
Conservation Purpose	0.411 (0.450)		
Other Purposes (Besides, Prevention and Conservation)	1.575** (0.613)		
Purpose of good entered as binary variables			
Conservation Purpose	-0.095 (0.427)	-9.996*** (0.821)	
Other Purposes	1.331*** (0.464)	-0.731(0.669)	

Variables	**Univariate meta-regressions**	**Multiple meta-regressions**	
	Coefficient (SE)	**Base Coef. (SE)**	**Reduced Coef. (SE)**
Prevention Purpose	-0.694* (0.325)		
Duration between surveys			
Hypothetical and Actual surveys held concurrently	0.930** (0.387)	-0.503 (0.381)	-0.497*(0.270)
Payment duration			
One-off payment elicited	3.364* (1.707)	-	-3.107**(1.262)
Payment vehicle			
Cash fee payment vehicle	0.174 (0.436)	-3.856*** (1.406)	-2.829*(1.502)
Type of comparison			
Between sample comparisons	-1.411* (0.800)	-0.393 (0.614)	
Survey setting			
Surveys held in a field setting	-1.029** (0.387)	-	
Money effects			
Money given for participation in hypothetical survey	0.333 (0.388)		
Money given for participation in actual survey	0.908** (0.347)	-0.069 (0.485)	
Comparisons between study attributes in hypothetical and actual surveys			
Sample type [Student or not]			
Student sample in both surveys	1.029** (0.387)	5.1999***(1.405)	5.192***(1.158)
Sample type [Potential user or not]			
Respondent a potential user in both surveys	0.864*** (0.222)	0.726 (0.617)	
Sample selection			
Same sample selection method in both surveys	0.737*** (0.242)		
Random sampling in both	-1.482** (0.589)	4.630***(1.069))	4.721***(1.176)
Purposive sampling in both	0.976** (0.430)	4.500*** (0.425)	4.441***(1.052)
Convenience sampling in both	1.029** (0.387)	-	
Administration mode			
Same mode of administration in both surveys	3.364* (1.707)		
Mail administration in both surveys	0.813** (0.376)	-4.287*** (1.445)	
In-person surveys in both surveys	0.095 (0.427)	-	12.53***(2.442)
Elicitation method			
Same elicitation method in both surveys	-0.527 (1.028)		
Bidding method in both surveys	-0.782*** (0.241)	-5.198*** (1.089)	-4.786***(1.151)
Dichotomous choice methods in both surveys	-0.447 (0.418)	-5.127*** (1.087)	-4.478***(1.010)

Variables	**Univariate meta-regressions**	**Multiple meta-regressions**	
	Coefficient (SE)	**Base Coef. (SE)**	**Reduced Coef. (SE)**
Open ended methods in both surveys	0.936** (0.415)	-4.724*** (1.558)	-4.009***(1.048)
Observations		56	56
R-squared		0.882	0.866

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

6.4 Discussion

The analyses presented in this chapter involve the synthesis and comparisons of multiple estimates from a broad range of goods and services. The analyses build on the systematic review, narrative summary and initial quantitative analysis presented in chapter 5. The current analysis is based on a larger dataset than previous meta-analysis. The degree of variation in the hypothetical and actual surveys summarised as ratios in chapter 5 provided estimates of 3.2 for mean ratios and 5.7 for percentage summaries. By pooling the values in the random effects meta-analyses, the magnitude of hypothetical bias for both the mean and percentage summaries is close; 1.7 for mean summaries and 2.3 for percentage summaries. However, as expected, the degree of heterogeneity in both analyses is very high at 99.6% for mean summaries and 90.2% for percentage summaries.

Using the expanded dataset, the magnitude of hypothetical bias identified in this analysis is very close to the estimate provided by the last meta-analysis (2.6) conducted more than a decade ago (Murphy et al. 2004). While this result adds credence to previous estimates of the magnitude of hypothetical bias for the range of goods / services investigated, the current analysis suffers from the problem faced by other analysis. The incidence of missing data and incomplete information on the estimation of WTP for a broad range of attributes in majority of criterion validity studies is high. This hinders a more robust analysis of the factors influencing hypothetical bias.

Investigation of the factors which explain hypothetical bias leads to mixed results, for most of the variables. This conclusion is similar to previous meta-analyses. For the mean summaries, robust results are found for the WTP elicitation method, duration of payment, the sector within which the valuation good falls, the purpose of the valuation good, sample selection models and the survey administration modes. For all the models (univariate, base and reduced models), hypothetical bias is higher when open ended methods are used to elicit WTP values and when one-off payments are elicited. Similarly, across the three models, hypothetical bias is lower when the valuation good is from the health sector, with prevention (and other – non-conservation) goods, when random sampling methods are used in both hypothetical and actual surveys and when in-person surveys are used.

Robust coefficients for the percent summaries include the income level of the country within which the study was conducted, the sample type and the WTP elicitation methods. In this, hypothetical bias is higher for studies conducted in high income countries and with the use of student samples. On contrary, hypothetical bias is lower when bidding and dichotomous choice methods are used in hypothetical and actual surveys of WTP. However, none of the robust regressions are similar across the mean and percentage summaries.

Like in the previous meta-analyses (Murphy et al. 2005), the use of student samples has been associated with higher hypothetical bias, but only with the percentage summaries. Further, the choice of elicitation methods is associated with hypothetical bias. This result is similar to previous meta-analyses (Murphy et al. 2004; List & Gallet 2001). The current analyses is the first to investigate the differences in hypothetical bias as a result of the sector within which the valuation good falls, the country within which the study falls, the purpose of the good, duration between the hypothetical and actual surveys, duration of payment, study administration mode and the sample selection method for both the hypothetical and actual surveys. The results of three of these variables are robust across all regression models as discussed above. The results for the remaining variables are not conclusive with this analysis.

Mixed results are obtained from several of the variables, with coefficient signs and directions changing across the univariate and both base and reduced meta-regressions. With this, it is not possible to make a conclusion on the effect of these variables on hypothetical bias. In addition, some of the variables that have been identified as influencing hypothetical bias in previous meta-analyses have not been confirmed in the current one. For example, while List and Gallet (2001) identified the class of the valuation good as influencing hypothetical bias, this variable was not significant in this analysis (List & Gallet 2001). As with the List and Gallet meta-analysis (2001) the current analysis finds no firm conclusions on effect of survey setting (laboratory or field) and comparison type (within or between) on hypothetical bias. Further, the review generated very few WTA comparisons that met the study inclusion criteria (n=4) and so the effect of the technique on hypothetical bias was not assessed.

This analysis aimed at updating the evidence on hypothetical bias, based on the current literature on criterion validity assessments. This further build on the evidence on the influence of certain experimental protocol on hypothetical bias. As the choice of predictors is largely exploratory rather than building on any theory, the results may be attributed to the model specifications.

The magnitude of hypothetical bias evidenced in this analysis and the most recent one by Murphy et al (2004) are both similar and amenable to the calibration suggestions by the NOAA panel (Arrow et al. 1993). Further, particularly for the percent summaries, the results support the recommendation by the panel on the use of dichotomous choice elicitation methods in efforts to reduce hypothetical bias. The evidence on the higher hypothetical bias with open ended methods further supports this recommendation. These results, while building on the earlier meta-analysis, are still early efforts to identify and quantify the magnitude of hypothetical bias in hypothetical and actual surveys of WTP. As suggested in the previous chapter, future efforts will be supported by the development of guidelines for the conduct and reporting of CV WTP criterion validity assessments

While the analysis presented in this chapter offers some promising results on the magnitude of hypothetical bias, some limitations may have affected both the analysis and the interpretation. The meta-analysis did not use all the estimates from the systematic review of criterion validity. As discussed, some of the estimates reported for the hypothetical WTP surveys could not be compared to estimates reported for the actual surveys. The results and conclusions on the presence of hypothetical bias in CV-WTP studies might have been different if the entire dataset was used. However, more than four-fifths of the dataset was used and the estimates from the remaining one-fifth are unlikely to change the conclusions significantly. Further, the estimates were included in the analysis as they were reported by the study authors. Adjustments of the estimates for inclusion in the analysis would have possibly introduced reviewer interpretation biases and potential errors.

The analysis was also limited to the reported variables and estimates. There was variation in the reporting of criterion validity studies coupled with missing and incomplete data on some key design attributes and respondent characteristics. This therefore limited the range of analysis that could be conducted. Further, in the absence of a theoretical framework to guide the specification of a model to investigate hypothetical bias, an exploratory approach was adopted. It is possible that alternate specifications of the regression models would have yielded different results. However, model diagnostics indicated that these were well fit for the available data. Further, to ensure credible interpretations of the results, evidence from previous reviews was adopted and this confirmed the majority of the expectations.

To permit the inclusion of all the identified studies in the quantitative synthesis, I could have contacted the study authors requesting for the datasets to extract the missing data for this synthesis. However, I did not think that the additional estimates would alter the current findings significantly and therefore this approach was not explored.

6.5 Implications for criterion validity

The presence of and magnitude of hypothetical bias remains a key issue affecting the large-scale use and acceptance of CV-WTP estimates across sectors. As shown in the review and meta-analysis, the method is least used in the health sector, even though results from studies in this sector point to lower hypothetical bias, compared to the other sectors. The magnitude of hypothetical bias found in the meta-analysis is also close to the estimate found in the earlier meta-analysis, even with the larger dataset.

While these results are promising, the limited empirical assessments of criterion validity present a challenge in several ways. Firstly, with even fewer studies identified from sectors such as health, generalizing the results presented in this chapter would be inaccurate. Secondly, the investigation of the different experimental protocol which might influence hypothetical bias is limited. Previous reviews and meta-analyses identified study elements such as the WTP elicitation technique, as being possible drivers of hypothetical bias. However, the results from those reviews are mixed, calling for additional research. Third, study results are reported variedly, with missing data in some studies hindering the synthesis of such for a pooled estimate of hypothetical bias. Of greater concern is the methods used to assess criterion validity. The limited empirical assessments especially in the health sector do not allow for the interrogation of the methods used, and hence the conclusions thereof. As a result, the current views on the criterion validity of CV-WTP estimates warrant further investigation.

6.6 Conclusions and chapter summary

While the presence of hypothetical bias in CV-WTP studies is acknowledged, the assessment of criterion validity is limited. This is especially so in the health sector.

To contribute to the evidence base on the criterion validity, further empirical assessments of criterion validity are necessary. Investigations should explore the effect of analytical methods on the assessment of and conclusions on criterion validity. Based on the findings of the current review and meta-analysis, criterion validity assessments are based on comparisons of summary estimates of hypothetical WTP and actual values. Often, an aggregate estimate from the hypothetical survey is presented as a comparator, even when multiple elicitation methods are utilised. An empirical study in which hypothetical WTP is elicited using multiple formats or at different levels would be best suited for such analyses. Using such a dataset, the alternate analysis of CV-WTP data elicited using the multiple methods and/ or elicitation points would be demonstrated. The results would then be used to illustrate the effect of analytical methods on criterion validity assessments and conclusions.

In further advancing the aims of this thesis, additional empirical analyses address this objective. The analysis of hypothetical WTP data significantly affects the assessment of criterion validity, and conclusions thereof. In the next chapter, a suitable dataset for such analysis is considered.

Chapter 7 Data

The systematic reviews and meta-analyses presented in chapters 4 and 5 identified significant differences in the conduct of CV studies and the analysis of estimates derived thereof. In chapter 6, the magnitude of the hypothetical bias was quantified. The key drivers of hypothetical bias were also identified. A key conclusion from this meta-analysis is that the magnitude of hypothetical bias in criterion validity assessments may not be as significant as authors report. An issue of primary concern, which may explain hypothetical bias, is a consideration of the study designs adopted to derive WTP estimates. In particular, the analysis of hypothetical WTP data significantly influences conclusions on criterion validity. This is especially critical where multiple elicitation techniques are used. As was discussed in chapter 3, the analysis of WTP data is driven by the choice of the WTP elicitation format. In chapter 5, it was established that studies often present a single summary statistic (for instance the mean) from the analysis of hypothetical WTP data, for comparison with actual survey values. Based on this comparison, assessments and conclusions on criterion validity are made. It is likely that WTP estimates, even from the same sample, will be different, depending on the elicitation technique employed. Conclusions on criterion validity made based on such comparisons might therefore be incorrect.

To further explore the effect of multiple elicitation methods on the assessment of criterion validity, such analysis are considered. A suitable empirical dataset was sought to address the aims of the thesis. In this chapter, the details of the dataset search and results are presented. The chapter is structured as follows: in section 7.1, the methods used in the determination of a suitable dataset for the planned analyses are discussed. The chosen dataset is presented, and a justification made for it in section 7.2. In section 7.3 the independent variables that will be used in subsequent analysis are justified and the chapter is summarised in section 7.4.

7.1 Justification for the choice of the dataset

7.1.1 Considerations for a suitable dataset

The systematic review of criterion validity assessments presented in chapter 5 and the subsequent meta-analysis in chapter 6 identified only twelve studies conducted in the health sector. Studies included in both the review and the meta-analyses:

1. Used direct stated preference techniques to elicit WTP values;
2. Conducted both hypothetical and actual surveys, with an actual transaction included and;
3. Provided empirical estimates of WTP/WTA.

These three criteria offer an opportunity for an analysis of criterion validity with minimal biases and ambiguity in the estimates used. As has been discussed in earlier chapters, carefully designed and executed stated preference studies can enable accurate estimations of the economic value of non-marketed goods. Given the hypothetical nature of CV studies and the nature of the valuation goods, it is not always possible to correctly estimate the actual values placed on goods. However, decision makers need to utilise information from the CV studies for evaluations such as CBA and for pricing decisions. To establish criterion validity, a gold standard is required. For CV studies, simulated market experiments (SMEs) based on actual cash transactions provide a validity criterion for comparison with hypothetical values (Champ et al. 2017). Nevertheless, a well-constructed SME in itself does not address the concerns with the hypothetical CV study.

A key consideration in the design of CV studies is to make the scenario and study as believable as possible (Tussupova et al. 2015; Andersson & Svensson 2008). One of the ways of ensuring this is by using elicitation methods that are familiar with respondents. These include dichotomous choice elicitation methods whose strength has been discussed elaborately in chapter 3. Dichotomous choice questions closely mimic real market transactions. Further, the use of DC methods in the conduct of CV studies has been widely recommended since it was popularised by Hanemann (Hanemann 1984). The method obtained further stamping with the NOAA panel (Arrow et al. 1993) recommending its use.

In addition to the elicitation questions, it has been suggested that CV studies collect a range of socio-economic attributes from respondents (Arrow et al. 1993). A minimum set of these attributes might include the respondents' age, education level, respondent's occupation, household size and composition and household income and expenditure. The analysis of these helps in identifying the factors which influence WTP and this can be used to assess the theoretical validity of the estimates obtained (Bateman et al. 2002). In addition, the results can be used to further refine the study tools.

7.1.2 Criteria for selecting a suitable dataset

To address the objectives of the thesis, the following criteria were used to determine a suitable dataset for the planned analysis. The CV study should:

1. Relate to a health good or service;
2. Elicit WTP using multiple methods in the hypothetical survey;
3. Include both hypothetical surveys of WTP and a SME (with respondents informed during the hypothetical elicitation process that SME involving actual cash transactions would be conducted. This enhances consequentiality).
4. Collect a minimum range of respondent SES variables to aid in the determination of the theoretical validity of the study and an exploration of factors expected to influence WTP.
5. Provide a sufficient description of the study process, such as the hypothetical scenario setting and the processes of determining the value cues provided. This will aid in the assessment of the content validity of the study tools.
6. Availability of the dataset. The authors or investigators of studies would be contacted with a request for the datasets.

The processes used to search and determine the most suitable database for the analyses are discussed in the next sections.

7.1.3 Data search process and results

A primary study was not conducted primarily due to time and resource constraints. Further, the use of secondary datasets has been lauded as a cost effective way of utilising all available data (Cheng & Phillips 2014; Vartanian 2011). In addition, as the thesis focusses on methodological challenges, a primary dataset was deemed not to be the only option. Therefore, a search for suitable secondary datasets was conducted from December 2015 to February 2016. The following process was pursued:

1. Suitable datasets were identified from the systematic review presented in chapter 5.
2. The twelve health studies identified from the systematic review of empirical studies assessing the criterion validity of CV-WTP were assessed using the criteria outlined in section 7.1.2. Two of these satisfied the criteria (see table 7.1 for the assessment outcomes).
3. Following discussions with experts in the field, including an investigator of one of the two identified studies, a decision was made to use one of the datasets, which was locally available (Bhatia & Fox-Rushby 2003). This dataset is discussed in the following section.

Table 7-1: Assessment of potential datasets for criterion validity analysis

No.	Study Ref	Multiple Elicitation methods	Hypothetical Scenario described	Determination of value cues discussed*	Respondents aware of actual study	Report SES variables[23]	Availability #~	Potential Dataset
1	(Bhatia & Fox-Rushby 2003)	Yes	Yes	Yes	Yes	Yes	Yes	Yes
2	(Blumenschein et al. 2008)	No	Yes	Yes	Yes	Yes	~	No
3	(Blumenschein et al. 2001)	No	Yes	No	Yes	Yes	~	No
4	(Bratt 2010)	No	Yes	Yes	Unclear	Yes	~	No
5	(Bryan & Jowett 2010)	Yes	Yes	No	Unclear	Yes	~	No
6	(Fox et al. 1998)	No	Yes	No	Yes	Yes	~	No
7	(Onwujekwe et al. 2002)	Yes	Yes	No	Yes	Yes	~	No
8	(Onwujekwe & Uzochukwu 2004)	Yes	Yes	No	Yes	Yes	~	No
9	(Onwujekwe 2004)	Yes	Yes	Yes	Yes	Yes	Possibly	Yes
10	(Onwujekwe et al. 2004)	Yes	Yes	No	Yes	Yes	~	No
11	(Onwujekwe 2001)	Yes	Yes	No	Yes	Yes	~	No
12	(Vernazza et al. 2015)	No	Yes	No	Yes	Yes	~	No

* where necessary depending on elicitation method # focus was on the hypothetical survey dataset ~ Not explored further

[23] E.g. respondents' age, education level, respondent's occupation, household size and composition and household income and expenditure

7.2: The Malaria WTP study

The study context within which the malaria WTP study was conducted is detailed in Appendix 19. The study is discussed more elaborately in the following sections.

7.2.1: Study aim

An economic evaluation was conducted to estimate the cost-effectiveness of malaria control strategies including In-house residual spraying (IRS) using deltamethrin, Deltamethrin[24] treated mosquito nets (TMNs) and Active case detection and treatment (ACDT). In addition, household's willingness to pay (WTP) for ITNs was estimated (Bhatia 2000). The evaluation was part of a large randomised controlled trial (RCT) in Surat, India, examining alternative approaches for controlling malaria. The study was conducted as part of the Malaria Control Research Project (MCRP) activities in India. The mandate of the MCRP in India was to reduce the impact of malaria in the population and through the research activities, contribute to the development of a national malaria control policy.

Intervention groups and comparison

The community RCT consisted of three groups covering the three main malariogenic zones of Surat district, to compare the effectiveness, efficiency and acceptability of malaria control interventions. The interventions tested were:

(a) Deltamethrin treated mosquito nets (TMNs);

(b) In-house residual spraying (IRS) using deltamethrin.

(c) Active case detection and treatment (ACDT), which was also the control group for all study arms. The effectiveness of the TMNs and IRS was compared against the control group.

(d) A fourth cluster was determined outside the trial area (OTA) where it was determined that the effects of the ongoing interventions would not be experienced by the households. This cluster was approached for the SME following the hypothetical survey.

[24] Deltamethrin is a pyrethroid insecticide that kills insects on contact and through digestion(Worthing 1983)

7.2.2: Study methods

Sampling and sample size

The unit of randomization in the study was the village. Within the village, a household was taken as the sampling unit. This is because in this setting, households were found to be the decision-making unit. Further, the health of one member affects other family members. A total of forty-two clusters, each consisting of 3 proximal villages were formed for the intervention and control arms of this study. For the OTA sample, one village was formed for purposes of this study. The final sample comprised of 300 households each from the 4 villages, for a total of 1,200 households (Bhatia & Fox-Rushby 2002).

Study procedures

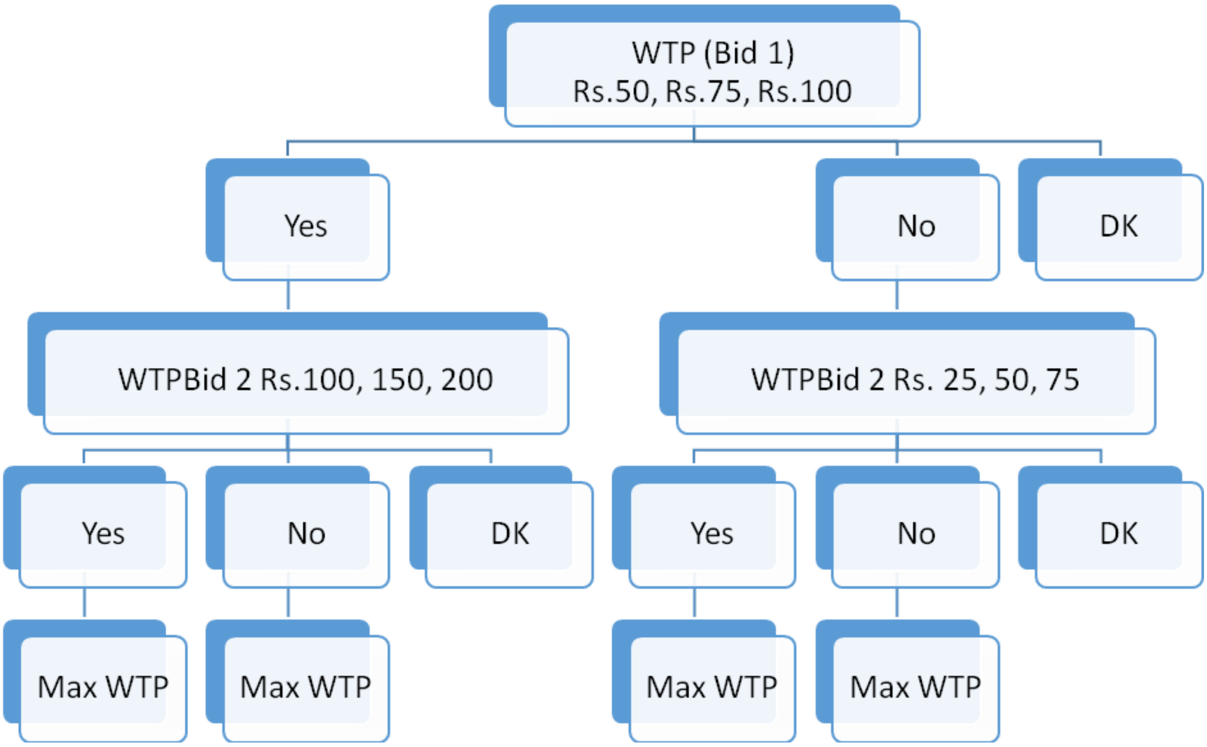
In the TMN villages, the required number of deltamethrin treated nets per household was distributed for free. This was estimated at 0.6 for each member of the household. Houses in the IRS villages were sprayed with deltamethrin at no cost to the household. In the active case detection (ACDT) group, early diagnosis for malaria was done through bi-weekly project worker visits to households. Cases that were detected during the visits were treated. There was no intervention in the outside trial area villages (Misra 1999).

7.2.3: Hypothetical survey

In addition to exploring the cost-effectiveness of the strategies, the study also estimated household's WTP for the malaria control interventions: treated mosquito nets (TMNs) and in-house residual spraying (IRS). Respondent's WTP was elicited through a two-stage bidding game with an open-ended final question. The respondents who were asked the valuation questions had been identified earlier through a question which asked whether they would be willing to buy TMNs either in cash or through instalments. Respondents who were not willing to buy TMNs either in cash or through instalments were assumed not to be in the market for TMNs and were therefore not subjected to the elicitation process. Three pre-determined starting bids (Rs.50, Rs.75 and Rs.100) were randomly allocated to respondents who were deemed to be in the market for TMNs (See figure 7.1 for the possible bid paths given these starting bids). The respondents were offered three choices: Yes, No and a "don't know" choice. The third choice (don't know) was included to ensure that respondents were not forced into giving a Yes or No response, as recommended in the literature (Harris et al. 1989; Arrow et al. 1993). The bids were increased or decreased once, depending on the responses to the first bid. After the two bids, all respondents were asked, in an open-ended question format, the maximum amounts that they were willing to pay for one TMN.

In addition to the questions about their preferences, socioeconomic and demographic characteristics of the households were documented. Information was also sought about the households' experiences with malaria, treatment seeking behaviours and the use of preventive measures as well as their income and expenditure patterns. Households in the four study arms were visited for up to three times in efforts to conduct the interview with the person regarded as the main earner. When this failed, another adult member of the household was interviewed for this study. An extract of the hypothetical survey questionnaire is provided in (appendix 20).

Figure 7-1: Bidding format used for elicitation of WTP values



Adapted from (Bhatia, 2000)

7.2.4: Simulated Market Experiment (Actual survey)

To check whether people's stated preferences in the hypothetical survey match their actions with regard to TMNs, a simulated market experiment (SME, also referred to in the document as "the actual survey") was conducted within a period of two months following the first survey. In such a survey, respondents provide actual values and are expected to make payments for the valuation good or service, taking ownership at the end of the process. Three hundred households (300) in the 20 villages outside the trial area were visited for a follow-up visit during which TMNs were sold to them at a fixed price. The median price of Rs.75 for one TMN from the hypothetical survey was given to the respondents in the SME in a dichotomous choice question format. Bhatia and Fox-Rushby (2002) details the design of this WTP study and the findings of related analysis provide some background information and depth that the current analysis builds on (Bhatia & Fox-Rushby 2002; Bhatia & Fox-Rushby 2003).

7.2.5: Ethical Considerations

The WTP study was part of a larger "Malaria control and research project" developed under Indo-UK technical cooperation and funded by the Department for International Development (DfID) UK and the Government of India (GoI). The project, being multi-disciplinary, also supported economic research within the study area, hence the cost effectiveness study through which the WTP survey was conducted. The principal investigators in the studies obtained ethical clearance from the institutions and governments involved, namely the University of London, the UK's Department for International Development (DFID) and the relevant ethics bodies within the governments of India and UK, including the programme in India, National Malaria Control Programme (NMCP) (Misra 1999; Bhatia 2000).

The planned analysis is therefore based on the secondary data as no additional primary data has been collected. Permission to use the dataset was obtained from the primary co-investigator and the dataset was availed by a second co-investigator in the WTP study and a custodian of the dataset. The anonymised dataset was made available as a stata file. In addition to the data, the hypothetical survey study questionnaire and coding frame were also provided. Research integrity was reviewed by the Brunel University Ethics team and approval granted (see Appendix 21).

7.3 Study variables

In previous chapters, it was suggested that hypothetical bias may be influenced by the choice of the elicitation method in hypothetical surveys of WTP and the analysis of the data thereof. Further, as discussed in the systematic reviews (chapter 5), studies do not report a range of variables which might influence hypothetical bias. The investigation of these variables might help in the interpretation of criterion validity assessments. However, in the absence of guidelines for the conduct of criterion validity assessments in CV WTP studies, one might ask what the most appropriate variables to investigate might be. As discussed in chapter 5, previous systematic reviews of criterion validity assessments tested a range of variables in an exploratory or hypothesis building manner. The results of these were largely mixed. In the next section, the variables in the Malaria dataset relevant for the current analyses are briefly outlined. This is followed by a discussion of the choice and justification of the independent variables used in subsequent analyses.

7.3.1 Variables in the malaria dataset

The Malaria WTP study contains over 200 variables. However, for this analysis, only those variables related to willingness to pay for TMN elicited using the bidding format and open-ended methods are used. These include the independent variables clustered below and detailed in table 7.2.

1) Respondent and household characteristics (gender, religion, caste, type of house, household size and number of children in the household, main earner in the household);
2) Socio-economic characteristics (education, occupation and income);
3) Malaria variables (knowledge, exposure and experience with the disease, perception of mosquitoes, knowledge of and use of prevention methods);
4) Treated malaria net variables (current ownership and source of current net).

Table 7-2: Variables in the malaria WTP Study

Variable	Variable type	Variable	Variable type
Household background characteristics			
Intervention group	Categorical	Type of house	Categorical
Household size	Continuous	Survey respondent [main earner or not]	Binary
Number of children aged less than 6 years	Continuous	Education qualification [Main earner]	Categorical
Sex [Main income earner]	Binary	Respondent Occupation	Categorical
Religion	Binary	Total household annual income	Continuous
Caste	Binary		
Malaria preventive measures			
Mosquito nuisance	Categorical	Whether the preferred method is a net or not	Binary
Total number of mosquito prevention methods known	Continuous	Total number of mosquito prevention methods used by respondent	Continuous
Net ownership variables			
Household net ownership	Binary	Number of nets purchased through market	Continuous
Number of nets owned	Continuous		
Malaria knowledge and experience variables			
Disease caused by mosquitos	Binary	Expenditure incurred on treatment	Continuous
Family members suffering from malaria last month	Continuous		

7.3.2 Choice and justification of independent variables

To avoid mis-specified models, the choice of the variables used in planned analyses was informed by theory on the subject and empirical evidence on WTP for treated mosquito nets. While some evidence (strong, weak or mixed) exists for some of the variables, there is no evidence for others. Theoretical justifications are provided for additional variables. Variables for which there is no evidence on their effect on stated WTP values are included in the models in an exploratory manner. Finally, some variables which have not been investigated in previous literature are included in this analysis as hypothesis testing variables. In the analysis, where variables signs are as predicted by existing literature will be an indicator of a good specification and vice versa. A review of studies assessing WTP for malaria nets was conducted to determine the evidence for the different independent variables expected to influence WTP. In the next section, the methods and results of this review are presented. A general discussion on the variables based on economic theory and the exploratory analysis concludes the section.

7.3.3 Systematic review of studies assessing WTP for treated mosquito nets

The purpose of this review was to summarise the evidence on variables which influence WTP values for mosquito nets. In addition, the review sought to determine the direction and strength of the variables and justify the inclusion of these variables in the analyses presented in this thesis. The review search methods and results are presented below.

7.3.3.1 Review search methods

The review was guided by the PRISMA statement on the conduct of systematic reviews (Moher et al. 2009). The review strategy was informed by earlier systematic reviews and meta-analysis conducted on WTP for malaria control interventions (Trapero-Bertran et al. 2013; Kutluay et al. 2015). The Scopus[25] database was searched from inception to January 2017. An advanced google search was also conducted to identify grey literature. In addition, reference lists of key papers were scanned for additional papers. The following net related terms (a) “treated net”; (b)

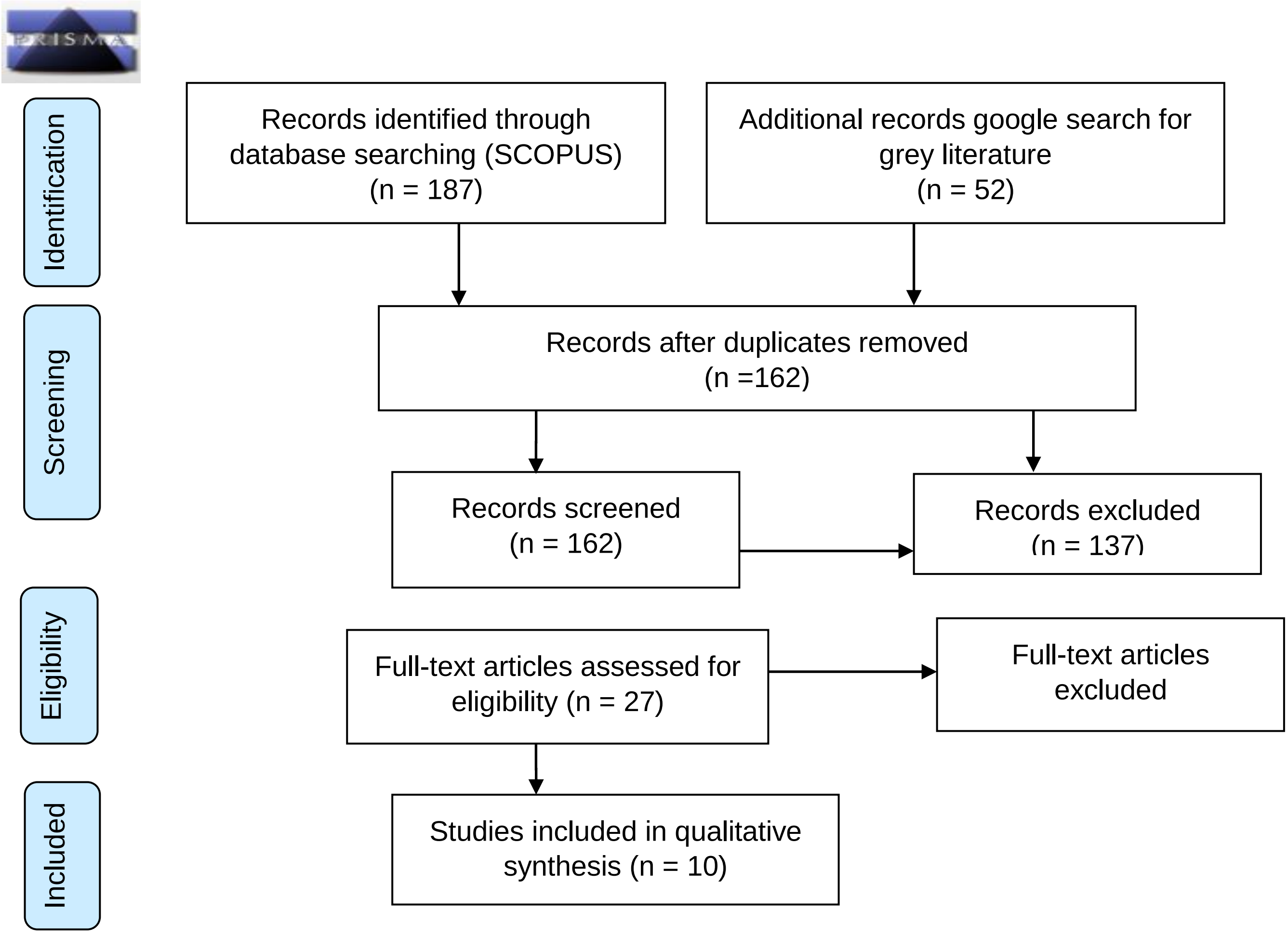
[25] Scopus is the largest abstract and citation database of peer-reviewed literature and was deemed sufficient for purposes of this review. Final articles were checked against previous systematic reviews on WTP for malaria prevention methods (Trapero-Bertran et al. 2013; Kutluay et al. 2015) to check that all relevant articles had been captured and this was confirmed.

"mosquito net"; (c) "bed net"; (d) "insecticide net" were crossed with the valuation terms (i) "willing" or "willingness to pay", or "willingness to accept"; (ii) "contingent and value or valuation"; (ii) "benefit" and "value" or valuation". Papers were included if they were: (1) published in English (2) conducted stated preference studies and not observation or other type of studies and excluded otherwise.

7.3.3.2 Search results

The PRISMA diagram presented in figure 7.2 outlines the search process and flow of records. From a total of 239 records identified from the search, data was extracted from 10 articles which satisfied the inclusion criteria. As the focus of this was primarily on the variables expected to influence WTP for treated mosquito nets, only these results will be presented. The list of included studies is presented in appendix 22.

Figure 7-2: Flow of articles for systematic review of studies assessing WTP for treated mosquito nets


PRISMA
Identification
Screening
Eligibility
Included
Records identified through database searching (SCOPUS) (n = 187)
Additional records google search for grey literature (n = 52)
Records after duplicates removed (n =162)
Records screened (n = 162)
Records excluded (n = 137)
Full-text articles assessed for eligibility (n = 27)
Full-text articles excluded
Studies included in qualitative synthesis (n = 10)

7.3.3.3 Evidence on variables influencing WTP

Detailed descriptions of the variables which influence WTP for TMNs as identified from the reviews are presented in table 7.3. Further, a summary of the expected signs for the final variables used in the analysis and the expected signs from the evidence from the reviews as well as economic and other exploratory variables is presented in table 7.4. The evidence on three key variable clusters which are most investigated in the literature is further discussed in the following section. Theoretical expectations for these variables and others for which there was no empirical evidence are discussed too.

Respondent and household characteristics

The evidence on the relationship between household size and WTP in the published literature is mixed. Onwujekwe et al (2002), in a study aimed at determining altruistic WTP for ITNs established a positive correlation with the number of people living in a household (Onwujekwe et al. 2002). Still, in a different study, respondents from households with many residents was associated with higher WTP for ITNs (Onwujekwe et al. 2004). A study by Biadgilign, Reda and Kedir (2015) established that the relationship between WTP and household size was not significant (Biadgilign et al. 2015) while another study conducted in Ethiopia established that there was a decline in the maximum willingness to pay amounts as the number of family members increases (Taye 2002). There is limited evidence on the effect of the household composition on stated WTP values for TMN. In a study by Chase et al (2009) the authors found that where a child under five is present households have a higher average WTP (Chase et al. 2009).

There is no evidence regarding the relationship between religion, caste and type of house on WTP for TMN. However, household caste[26] and the type of house can be used as proxies for wealth, given the current study setting. In this case, individuals from higher castes and those living in more permanent houses would be expected to be wealthier than those from lower castes and those living in the semi-permanent type of houses and might therefore provide higher WTP values. However, the same

[26] The different castes in the study setting are defined as footnotes to table 7.4.

group may not need as many TMNs as those living in semi-permanent structures as their housing provides more protection from mosquitoes, hence lower WTP values.

Socio-economic characteristics

Available evidence on the relationship between income and WTP for ITNs is mixed. While the estimation of income in these studies is varied, majority of the studies have established a positive relationship between income and WTP (Taye 2002; Aleme et al. 2014; Biadgilign et al. 2015; Onwujekwe et al. 2001). In these studies, WTP values increased with income. Further, the evidence on this variable is also supported by theoretical constructs which hold that, assuming that the TMN is a normal good, more nets will be bought (or higher WTP value stated), with an increase in income (Varian 2014). However, in one study conducted in Ethiopia, the researchers found that income had no effect on WTP for an ITN (Gebresilassie & Haile Mariam 2000). There is strong evidence too that higher starting bids lead to higher WTP values (Onwujekwe & Nwagbo 2002).

Malaria variables

For most of the variables related to knowledge and use of malaria prevention methods, there is no evidence on the expected relationship with WTP. Knowledge of the malaria prevention methods may lead to higher or lower WTP values depending on the respondents' preferences and whether they consider methods other than TMN as complements or substitutes. Where TMNs are regarded as compliments to other prevention methods the stated WTP values will be low while where TMNs are regarded as substitutes then the stated values will be higher. Net ownership may lead to lower WTP values as the households may not have a need for them.

There was no clear evidence on treated malaria net variables such as current ownership, source of current net and perception about number of nets owned by the household. These variables were therefore included in the models in an exploratory manner.

Table 7-3: Summary of factors influencing WTP for ITNs from reviewed studies

Source	Country	Respondent	Variable	Empirical evidence from paper
Gebresilassie and Mariam., 2000	Ethiopia	Household head or adult member of the household	Gender	Females were 0.47 times less likely to be willing to pay for ITNs than males; this was statistically significant (p=0.03) even after controlling for the possible confounders
			Education status	People who could only read and write were almost three times (p=0.000) and those who finished elementary school were 3.3 times (p=0.014) more likely to be willing to pay than illiterate ones.
			Source of bed nets	Families that obtained their nets by purchasing were 3.4 times more likely to be willing to buy than those that got it free of charge (p=0.000). The association became 2.34 times even after controlling for potential confounders (p=0.01)
			Use of ITNs	Those who used their ITNs were about four times more likely to be willing to buy than those who did not (p=0.001)
			Income	Monthly income of the households was not a significant determinant of people's willingness-to-pay
Mujinja and Sauerborn, 2004	Tanzania	Household head or adult member of the household	Net prices	There were no statistically significant differences between males and females who were willing to pay for an ITN at different prices in the two study rounds (p>0.05)
			Gender (Mean maximum WTP amounts)	The mean maximum WTP difference between men and women in the two rounds was not statistically significant (p=0.08 in first round and p=0.89 in second round)
			Respondent's under five child suffered from malaria in the last three months	There was no significant association between willingness to pay for an ITN and having a child with recent history of malaria, in both males and females (p=0.08 for males and p=0.30 for females)
			Recent experience of malaria episodes	Among both males and females, there was an association between a recent experience with malaria episode and WTP (p=0.05, χ2=5.92 and p=0.02, χ2=8.1). Moreover, the association was stronger among females than males.
			Altruistic behaviour: Willing to pay for another person in the household	No significant difference between genders for WTP for any other member of the household, including children under five. Among makes, there was no statistically significant difference in WTP for an ITN between those who had an under-five child and those who did not (χ2=1.74; p=0.42) There was a statistically significant difference in WTP for any other person in the household among females who had under-five compared to those who did not (χ2=6.40; p=0.041)

Source	Country	Respondent	Variable	Empirical evidence from paper
Taye, B, 2002	Ethiopia		Income	Household income positively influences the willing to pay decisions.
			Ownership of oxen (as a proxy for wealth)	The higher the number of oxen owned the greater is the willingness to pay for bed nets.
			Presence of family members suffering from malaria infection	Households whose family members are suffering from malaria infection are highly willing to purchase insecticide-impregnated nets. The higher the numbers of ill family members the larger is the households' willingness to pay for a measure that help reduce the malaria infection incidence.
			Household expenditure on malaria prevention	Total cost incurred by households due to malaria is found to be a statistically significant factor positively influencing the WTP decisions.
			Age	The amount the household is willing to pay for a bed net may fall with the increase in the age of the family head. Younger household heads have more preference for modern means of health care goods and services than older household heads.
			Education level of the household head	The coefficient for education is found to be negative and significant, indicating a decline in WTP amounts with education level.
			Family size of the household	The statistically significant and negative coefficients of this variable show that the decline in the maximum willingness to pay amounts as the number of family members increases.
			Sex (gender)	The coefficient for the sex of the household head is found to be negative and insignificant, implying that gender has no role in influencing the amount of money households would be spending on the purchase of bed nets.
			Size of land holdings (as a proxy measure for wealth)	The size of land a household owns is observed to have no effect at all on the amount it is willing to pay for mosquito bed nets.
Aleme et al. 2014	Ethiopia	Household heads or their representatives	Gender	Females showed a higher willingness to pay for ITN than males (AOR=1.86, 95% CI=1.29 - 2.55)
			Marital status	Respondents who were married, widowed and divorced had higher willingness to pay for ITNs
			Education	Respondents who completed primary school had higher WTP for ITNs (AOR=-4.72, 95% CI=1.48 – 15.04)
			Income	As the average monthly income of respondents decreased, the WTP for ITNs had increased significantly (AOR=-22.44, 95% CI=12 – 41.34)

Source	Country	Respondent	Variable	Empirical evidence from paper
			Knowledge about malaria	Respondents who had poor knowledge about malaria had less WTP for ITNs than respondents who had knowledge (AOR=--0.68, 95% CI=0.47 – 0.98)
			Perceived benefit of ITNs	Respondents who showed low perceived benefit of ITNs had significantly lower WTP for ITNs than higher perceived benefit (AOR=-0.28, 95% CI=0.2 – 0.4)
			Perceived susceptibility of malaria	Respondents who showed low perceived susceptibility of malaria had significantly lower WTP for ITNs than higher perceived benefit (AOR=-0.64, 95% CI=0.44 – 0.93)
			Perceived severity of malaria	Respondents who showed low perceived severity of malaria had significantly lower WTP for ITNs than higher perceived benefit (AOR=-0.65, 95% CI=0.47 – 0.91)
Biadgilign et al. 2015	Ethiopia	Household head or an adult household member	Average monthly income	Average income more than 10.4USD per month has a statistically significant effect on WTP (p=0.045)
			Distance in minutes to the health facility	Living within a distance of 30 minutes to the health facility had a statistically significant effect on WTP (p=0.048)
			Age	Not significant
			Occupation	Not significant
			Marital status	Not significant
			Education	Not significant
			Family size	Not significant
			Know benefit of a mosquito net	Not significant
			Family member travel anywhere in the last one month	Not significant
			Malaria can lead to death of children	Not significant
			History of malaria in the last one year?	Not significant
			Perception of family risk of getting malaria?	Not significant
Mujinja, 2006	Tanzania	Household head	Age	Statistically significant negative impact (Lower WTP as age increased) (p=0.00)

Source	Country	Respondent	Variable	Empirical evidence from paper
		or other adult representative	Distance to the nearest heath facility	Statistically significant negative impact (Lower WTP as distance to the nearest health facility increased) (p=0.00) In logistic model, this coefficient had a statistically significant positive impact on the likelihood of a Yes response (p<0.05).
			Recent experience with malaria (3 months before interview)	In the generalized lest squares model, this coefficient was positively statistically significantly associated with mentioning a higher maximum WTP (p=0.00) In logistic model, this coefficient had a statistically significant negative impact on the likelihood of a Yes response.
			Knowledge about how malaria is transmitted	Positively statistically significantly associated with mentioning a higher maximum WTP (p=0.00)
			Self-reported health status	Positively statistically significantly associated with mentioning a higher maximum WTP (p=0.00)
			Self-assessment of being able to buy an ITN	Positively statistically significantly associated with mentioning a higher maximum WTP (p=0.00)
			Perceiving mosquitoes as a nuisance	Positively statistically significantly associated with mentioning a higher maximum WTP (p=0.00) In logistic model, this coefficient marginally predicted probability of giving an affirmative response to a WTP response (p=0.1)
			Having an untreated bed net	Positive marginal impact on both mentioning the maximum WTP (p=0.07) In logistic model, this coefficient had a statistically significant positive impact on the likelihood of a Yes response (p<0.05).
			Price of the ITN	A negative and highly statistically significant value (p<0.05) for the coefficient on the price of an ITN variable.
Chase C et al., 2009	Mozambique	Head of the household or a representative over the age of 18	Formal schooling	Stated WTP is significantly higher for younger educated respondents with formal schooling contributing an additional $0.80 to average WTP (p<0.001)
			SES Quintile	Movement between SES quintiles increases (p=0.002), although the effect itself is small
			Formal schooling and higher SES combined	Respondents with both formal schooling and higher SES scores show no measurable increase in WTP
			Use of alternate method	Respondents reporting the use of alternate methods state a lower average WTP and this is significant (p=0.078)
			Respondents that had received IRS	Respondents that has received IRS state a lower average WTP and this is significant (p=0.046).
			Education	Positive and significant on WTP averaging $0.84

Source	Country	Respondent	Variable	Empirical evidence from paper
			Knowing where nets are sold	Positive and significant
			Children under 5 in the household	Where a child under five is present households have a higher average WTP ($p=0.004$)
			Females	Females state less WTP on average
			Head of household	Heads of households state less WTP on average
Onwujekwe & Obinna, 2002	Nigeria	Household heads or their representatives	Sex	Statistically significant ($p<0.01$ in Orba with the BWFU method and $p<0.001$ in Mbano)
			Education	Statistically significant ($p<0.05$)
			Presence of malaria in the household	Statistically significant ($p<0.05$)
			Expenditure on school fees	Statistically significant ($p<0.01$)
			Average monthly treatment cost of malaria	Marginally significant ($p=0.066$)
			Age (in Mbano with the bidding method)	Statistically significant ($p<0.001$)
			Number of people living in the household	Statistically significant ($p<0.056$)
Onwujekwe et al., 2001	Nigeria	Household head or a representative	Main savings scheme	The more enhanced savings scheme a household had, the more willing they will be able to pay for altruism ($p=0.04$)
			Sex	Men are more likely to pay for altruism than women ($p=0.021$)
			Marital status	Single people are more likely to pay than married ones ($p=0.053$) More significant in the reduced model ($p=0.05$)
			Willingness to pay for own ITNs	Respondent's WTP for their own nets was positively correlated with altruistic WTP ($p=0.000$)
			Monthly household expenditure to treat malaria	Positively correlated with altruistic WTP ($p=0.031$)
			Number of people living in a household	Positively correlated with altruistic WTP ($p=0.058$)
Onwujekwe et al., 2004	Nigeria	Household head or their representatives	Actual incidence of malaria	Positively associated with stated WTP for ITNs ($p<0.01$) and actual purchase of ITNs ($p<0.05$)
			Stated WTP	Level of stated WTP positively associated with actual purchases $p=0.001$

Source	Country	Respondent	Variable	Empirical evidence from paper
			Sales distance	Living further away from the sales points for the nets was negatively and significantly associated with actual purchase of ITNs (p=0.01)
			Education	Presence of formal education was positively associated with ownership of untreated nets and stated WTP for ITNs (p<0.01)
			Sex	Interviewing a male was associated with higher stated WTP for ITNS (p<0.05)
			Status in the household	Head of a household was associated with higher stated WTP for ITNs (p<0.05)
			No of residents in household	Respondents from household with many residents associated with higher WTP for ITNs (p<0.05)
Onwujekwe and Nwagbo, 2002	Nigeria	Household head or their representatives	High starting point bid	Significantly correlated with WTP for small nets and with a negative sign (p<0.01)
			Medium starting point bid	No significant relationship
			Status in the household	No significant relationship
			Gender	Negatively and significantly correlated with WTP in WTP for large nets (p<0.01)
			Age	Negatively and significantly correlated with WTP in WTP for large nets(p<0.01)
			Education level	Positively and significantly correlated with WTP in WTP for large nets (p<0.05)
			Occupation	Positively and significantly correlated with WTP in WTP for large nets (p<0.01)
			Marital Status	

Table 7-4: Summary of independent variables included in the empirical analyses

Variable	**Reported signs (Evidence)**	**Expected signs (explanation)**
Respondent and Household characteristics		
Intervention village TMN IRS ACT OTA	None	- (Respondents in TMN village expected to be less willing to buy or pay for nets as they already have them)
Gender (Male)	+	+ (Patriarchal society, men likely to be decision makers and budget holders)
Religion Hindu Christian Buddhist Other	None	+ (For people professing the Hindu faith as they are strategically placed for employment opportunities in this society)
Caste Scheduled caste[27] Scheduled tribe[28] Other backward caste[29] Other caste[30]	None	+(For people in higher castes)
Type of house Kaccha[31] Semi pucca[32] Pucca	None	+ (For households living in houses made with high quality materials (such as Pucca)
Household size	Mixed	+ (Higher household income if more adults in a household)
Children below 5 yrs. *in the household*	+	+ (children below the age of 5 more susceptible to Malaria infection)

[27] Scheduled caste: the official name given in India to the lowest caste, considered 'untouchable' in orthodox Hindu scriptures and practice, officially regarded as socially disadvantaged.

[28] Scheduled tribe: an indigenous people who are descendants of the tribal communities who primarily lived in the forest regions.

[29] Other backward caste: comprises of natives who belong to the Sudra Varna or the Sudra (lower) caste, formerly considered untouchables.

[30] Other Castes: categories of people who were converted from Hinduism to other castes

[31] Kaccha: A temporary house made of light materials.

[32] Pucca: A permanent house made of strong materials.

Variable	**Reported signs (Evidence)**	**Expected signs (explanation)**
Main income earner (Yes)	+	+ (The main income earner is likely to be budget holder too)
Socio-economic characteristics		
Education No education or primary education Further education	Mixed	+ (Higher education expected to lead to more employment opportunities and/or higher salaries). Higher education also equated with a greater understanding of malaria and the prevention mechanisms
Occupation	Mixed	+ (For respondents in more regular and better earning occupations)
Household income (month/annual)	+	+ (Higher disposable income)
Malaria variables		
Consider mosquitoes as nuisance? (Major, Minor, No	None	+ (Respondents who consider mosquitoes to be a minor or major nuisance more likely to be WTB and WTP more for TMN compared to those who do not consider mosquitoes to be a nuisance
Knowledge of malaria (Yes)	+	+ (Greater understanding of the threat=higher WTP)
Exposure to malaria (Yes)	+	+ (Previous experience of the negative consequences of the disease=higher WTP)
Previous episode of malaria within household (Yes)	+	+(Previous experience of the negative consequences of the disease=higher WTP)
Malaria treatment related costs	+	+ (Higher costs increase WTP for TMN)
Knowledge of prevention methods	+	+ (Knowledge here does not equate to experience but knowledge = higher WTP)
Use of other prevention methods	None	Mixed (Depending on whether TMN is considered a substitute or complimentary, and the price of the net in relation to the prices of other prevention methods)
Treated mosquito net (TMN) variables		
Current ownership of a TMN (Yes)	Mixed	Mixed (Current owners may not have need for more but the experience with a TMN may increase the WTP for an additional one
Current use of TMN (Yes)	+	+ (Positive experience = Higher WTP)
Source of TMN Purchased through the market Freely distributed through the project	 + -	 Mixed (Depends on factors such as availability of a free TMN or the affordability of market TMNs) -(Having received a free project net may reduce WTP for a TMN)

7.3.3.4 Further discussion on variables based on economic theory

As an economic good, an ITN is a private good with public characteristics and is thus regarded as a quasi-public good. Two key characteristics of a quasi-public good are that it is semi non-rival which means that up to some point, extra persons using the good do not reduce the benefit of the use of the good by the primary owner or user and semi-non-excludable which means that it is often difficult to totally exclude anyone who did not contribute to the purchase of the good from enjoying the benefits. In the case of an ITN, extra people benefiting from the presence of an ITN in a sleeping space even if they are not sleeping under it do not reduce the benefits of the net to the owner or user. Further, it is not possible to exclude others from enjoying the benefits of a mosquito free environment when mosquitoes are killed by the insecticide impregnated in the ITN. The discussion below on variables expected to influence WTP for an ITN draws from economic theory.

When facing the task of making a decision, the consumer may act in two ways: given the prices and income, he decides in order to maximise the utility; or, given the prices and a certain level of utility, he decides in order to minimise the expenditure (Mas-Colell et al. 1995). Broadly, income and prices and their determinants are expected to influence an individual's demand[33] for a commodity such as an ITN.

Income

For all individuals, income is a finite resource which when combined with commodity prices determines a budget set. The budget set consists of all consumption bundles[34] that the consumer can afford at given prices and income (Varian 2014). A budget line is used to illustrate the set of bundles that just exhaust the consumer's income. For a normal good, an increase in income allows an individual to purchase more of it, while a decrease in income reduces the quantity of the good purchased, all other factors held constant. An ITN is a normal good which means that the demand for it increases with increases in income and vice versa, *ceteris paribus* (Varian, 2014). Two common value elicitation mechanisms are used to illustrate this. With a

[33]Demand refers to the quantity of a commodity that an individual is willing and able to buy during a given time period (Hardwick et al. 1986).

[34] The term consumption bundles refers to a complete list of all the commodities either goods or services from which the consumer makes his choice (Varian 2014).

dichotomous choice question, individuals at higher income levels are expected to be more willing to accept higher WTP amounts, *ceteris paribus*, and vice versa. Similarly, under the open-ended elicitation mechanism, individuals with higher incomes are expected to be willing to pay more for a unit of an ITN than those with lower incomes.

Employment

Broadly, employment is expected to generate income for an individual and therefore an individual's ability to engage in transactions. Applying the economic theory on income, being in employment suggests that an individual has disposable income for use in the purchase of ITNs, ceteris paribus. An employed individual would thus be expected to be more WTP for an ITN than an unemployed individual. Further, an employed individual would be expected to be WTP higher prices for an ITN than an unemployed individual. The different levels and types of employment (e.g. casual, skilled, unskilled) further implies differing levels of income and the economic theory that at higher levels of income individuals are expected to be WTP higher prices and purchase higher quantities of ITNs would apply.

Household size

The size of a household is expected to determine the level of disposable income available for purchase of commodities such as ITNs. Generally, holding all other factors constant, a larger household implies more expenses on what may be regarded as basic needs with less income available for the purchase of ITNs. A small household on the other hand would be expected to have enough to meet their basic needs and purchase ITNs, *ceteris paribus*.

Household composition

The dynamics within a household are expected to influence WTP for ITNs. A household composed of adults might have more income if they are engaged in income generating activities. The economic theory on the effect of income on demand for commodities therefore applies. At higher combined incomes, the household would be expected to be WTP more and purchase more units of ITNs and vice versa. On the other hand, if a household comprises of younger members, such as children, the income available for purchases would be limited if they all depended on only one breadwinner.

Further, studies show that children under five years and pregnant women are disproportionately affected by malaria. Hence, the need for ITNs in such households is higher. It would be expected that such households would purchase more ITNs, *ceteris paribus*. In a study by Chase et al (2009) the authors found that where a child under five is present households have a higher average WTP (Chase et al. 2009).

Decision making in the household

In many settings, economic decisions in a household are made by the purse holder, the individual who earns the income. The choices of the decision maker may not always reflect those of other members of the household, especially those affected by such decisions. For instance, in the case of WTP for an ITN, while a woman may be WTP a given price of purchase a certain quantity of ITNs, she may not have the power to influence purchase decisions. On the other hand, knowing the constraints of his income, the decision making may not be WTP for ITNs or may be willing to pay different prices.

7.4 Conclusions and chapter summary

In this chapter, the process used to identify a suitable dataset for the empirical analyses presented in this thesis was presented. The systematic reviews on criterion validity illustrate the scarcity of criterion validity assessments, especially in the health sector. The limited empirical work in this subject also hinders any extensive interrogations of current claims on the hypothetical bias in CV-WTP studies. The best option for such analysis would be a primary dataset. Primary data collection would have been more amenable to the exploration of a range of experimental protocol related to the conduct of CV-WTP and the effect of these on criterion validity assessments and conclusions. Such a dataset would have offered the opportunity to include specific variables of interest and test for more experimental protocol.

The use of secondary data limits the range of analysis that can be conducted to the available variables. However, researchers agree that the use of secondary data is a cost-effective way of utilising all the available data on a subject (Cheng & Phillips

2014; Vartanian 2011). Further, the available dataset was relatively big and collected within a large clinical trial. This meant that robust checks were in place to ensure the quality of data. An empirical study of this scale and quality within the duration of this PhD research is unlikely to have been as robust as the identified dataset.

The identified dataset is also relatively old, having been conducted more than a decade ago. However, as the focus of the present analyses is primarily an investigation of the methods, with no interest in the actual WTP values for the mosquito nets, the dataset is considered appropriate for the planned analyses. Further, empirical assessments in the field have remained few. Even fewer are empirical assessments of hypothetical WTP employing multiple elicitation formats and estimates. Finally, the choice of the dataset followed a systematic search, lending credibility to the process.

A recurring theme in discussions on criterion validity is the effect of the elicitation methods on hypothetical bias. In the identified dataset, hypothetical WTP was elicited using multiple elicitation formats. Further, for one of the formats, the bidding format, WTP was estimated at different points. In the next chapter, an analysis of such discrete data is conducted. The results are used to illustrate the possible effects of the WTP analysis methods on the choice of comparator and on criterion validity assessments and conclusions thereof.

Chapter 8 Estimating WTP Bid Functions from Discrete Choice Data

8.1 Introduction

As discussed in previous chapters, the analysis of hypothetical WTP data directly affects conclusions on criterion validity. The majority of criterion validity assessments have been conducted by comparing estimates from hypothetical studies of WTP with actual values obtained through different methods as discussed in chapter 3. The systematic review on criterion validity presented in chapter 5 established that the majority of the empirical assessments conducted across sectors have concluded that CV-WTP does not demonstrate criterion validity. These conclusions are made primarily based on comparisons of hypothetical WTP and actual survey values. However, criterion validity can also be determined by analysing the determinants of the WTP values to determine whether these accurately predict actual values. The WTP values can also be predicted from such analyses. Despite the potential strength of such analyses, criterion validity assessments based on such predictions are limited in the literature.

Further, as discussed in chapter 3, the analysis of hypothetical WTP data must be guided by the elicitation question (Donaldson et al. 1998; Kurth et al. 2004). This analysis affects the predicted WTP estimates, the inferences we make about the determinants of WTP and criterion validity conclusions thereof. Finally, the choice of analysis technique directly influences the summary estimates presented for criterion validity comparisons and hence the conclusions. The aim of this chapter is to demonstrate how the analysis of WTP values elicited using discrete choice elicitation methods affects conclusions on criterion validity. The remainder of the chapter is organised as follows: in section 8.2, the methods used in this analysis are presented, beginning with a brief discussion on the WTP elicitation technique as used in the Malaria WTP study. The results of the analysis are presented in section 8.3. A discussion is presented in section 8.4 while in 8.5 the chapter is summarised.

8.2 Methods

8.2.1 The bidding technique

The bidding technique used in the malaria WTP study involved elicitation of hypothetical WTP values at two levels. While the use of multiple bids has been shown to increase the efficiency of welfare estimates (Welsh & Poe 1998; Alberini et al. 2003; Vossler et al. 2003), concerns around the use of the method have been raised (Veronesi et al. 2011; Vossler et al. 2004). In addition to bid design effects on the WTP preferences, the analysis of such data is prone to misinterpretation (Vossler et al. 2003). Where multiple bids are used, it is possible that different factors influence the WTP estimates at the different levels. Therefore, analysis of separate bid paths is likely to generate different results from an analysis of combined bids (with the bid paths accounted for). Further, summary statistics from WTP values elicited using the bidding technique can be elicited in multiple ways: percentages (respondents above or below the bid level) or as means estimated from appropriate bid functions. The choice of summary statistic for comparisons with actual values directly affects criterion validity conclusions. As a result, inferences made from the analysis of such data directly affect conclusions on criterion validity and the subject therefore warrants further investigation.

In summary, concerns on criterion validity conclusions based on multiple bids, with a single summary estimate arise because:

1. The predictors of WTP may be different at different bid levels indicating different underlying distributions of WTP or lack of stability in respondent preferences. Researchers have argued that this is not unusual, but the complexity of trying to isolate the possible causes of these differences has been observed (Alberini et al, 2005);
2. These predictors might be statistically different;
3. The predicted WTP might be different, and statistically so, at the different bid levels, and
4. The different predictors and predicted WTP affect conclusions on criterion validity.

As discussed in chapter 5, even with multiple estimates from WTP data, and the above concerns, often the justification for the choice of estimate used in comparisons with actual values is not provided. These limits further interrogation of conclusions on hypothetical bias.

Using the empirical dataset presented in chapter 7, I estimate WTP bid functions for the single choice DC (willingness to buy nets) and the two bid levels (WTP for treated mosquito nets) separately. Bid functions are also estimated for the entire sample while controlling for the different bid paths. As discussed above, literature on the analysis of WTP data proposes the inclusion of the different bid paths in the models, as possible predictors. However, supporting literature on the appropriateness of the separate analysis of multiple bids is lacking. The different analyses are conducted to further illustrate the range of WTP estimates and estimators that could be obtained with the use of such multiple bidding elicitation techniques. Comparisons are made of the two sets of analyses with discussions highlighting the differences in the predictors and predictions of WTP, depending on the choice of analysis. The variables and specific analysis employed in this chapter are presented in the next section.

8.2.2 Independent variables

The data used in this analysis was discussed in the previous chapter (7). All the independent variables used in this analysis were justified in the same chapter too (section 7.3.1). These were classified into the below clusters:

1. Respondent and household characteristics (gender, religion, caste, type of house, household size and number of children in the household, main earner in the household);
2. Socio-economic characteristics (education, employment and income);
3. Malaria variables (knowledge, exposure and experience with the disease, knowledge of and use of prevention methods);
4. Treated malaria net variables (current ownership and source of current net).

Correlation coefficients were determined for all independent variables that were selected for inclusion in the regression models (appendix 33) with decisions made based on Mukaka (2012) presented in appendix 32 and discussed in section 6.2.2.

8.2.3 Dependent variables

Building on the discussion in section 8.2.1, the dependent variables are based on the multiple WTP elicitation points which are the subject of the analysis in this chapter. These represent the multiple stages at which willingness to pay preferences were elicited in the Malaria WTP study.

As discussed in section 7.2.3, all study respondents were subjected to an initial elicitation question which identified those who were willing to buy nets either in cash or instalments. Those who were determined to be in the market for the nets were subjected to a two-stage bidding process (see figure 7-1). In this chapter, the dependent variables are based on these four WTP elicitation points. These are:

1. Willingness to buy treated mosquito nets in cash or instalments
2. Willingness to pay for one treated mosquito net at the first bid
3. Willingness to pay for one treated mosquito net at the second higher bid
4. Willingness to pay for one treated mosquito net at the second lower bid

8.2.4 Data analysis

The analyses were conducted in the four stages:

(1) Descriptive analysis to summarise the data. Comparisons were done across the intervention groups and chi-square tests used to test for differences in key attributes;

(2) Bivariate analysis to determine the unadjusted relationship between each of the independent variables on dependent variables;

(3) Multivariate analysis to investigate the adjusted relationships between the independent variables on the dependent variables. Through this, the predictors for WTP are determined for the different elicitation methods and bid levels.

(4) Predictions of WTP from the analysis in (3) above.

8.2.4.1 Regression Models

As discussed in section 8.2.1, two sets of analyses are conducted: (1) different bid paths and (2) combined bids. All the dependent variables in the analyses presented in this chapter involve binary outcomes (yes or no). In analysing binary outcome data, one can use either logit or probit regression models (Greene 2003). While both can be used in the same way, the main difference between them is theoretical and relates to the distribution of the error term (Harrell 2016). The logit distribution assumes that the error term follows a logistic distribution while a probit distribution assumes that this follows a normal distribution (Jones 2007; Fernando 2011). The choice between the two models is a matter of convenience (Chen & Tsurumi 2010;

Greene 2003). Therefore, probit regression models are estimated in the analyses presented in this chapter.

Both base and reduced models were derived. The base model included all the independent variables included in the analysis. Stepwise regression was used to derive reduced models from the base model. In this, independent variables that were not statistically significant (with p≤ 0.05) were identified and removed. Categorical variables were dropped from the model if they were jointly insignificant (p > 0.05) and were not used to identify the selection model. As suggested in the literature, the wald test was used to test for the significance of variables before their removal from the model (Agresti 1990; Bursac et al. 2008; Baum 2006). All the statistical analyses were undertaken using STATA version 14 software. The specific equations for each of the dependent variables are presented below.

Willingness to buy treated mosquito nets

The following equation is estimated to model the factors that influence willingness to buy treated mosquito nets (WTBNETS).

$\Pr(WTB_i = 1|x) = p(WTB1_i^* > 0|x)$ where,

Pr denotes the probability and x represents the vector of regressors which are assumed to influence the decision to purchase the valuation good. The model assumes that the error terms are independent and normally distributed.

Willingness to pay for one TMN at the first bid

As discussed in chapter 7, during the bidding process, respondents were allocated to one of three starting bids (Rs.50, Rs.75, and Rs.100). This was a binary choice question with two possible outcomes for each bid: Yes (WTPBID1=1) if this first bid is accepted and No (WTPBID1=0) if the first bid is not accepted.

However, willingness to pay for one TMN, for the first bid is only observed for respondents who have a positive, non-zero WTP. Respondents who were not willing to buy a TMN are deemed to have zero WTP. The reasons for the zero WTP were not explored in the study. In these analyses, these responses are broadly classified as zero bids. These are identified in the previous filter question (dependent variable 1: willingness to buy nets) which determines those who are willing to buy TMNs.

It is possible that respondents who choose not to buy TMNs are systematically different from those who chose to buy TMNs and are therefore regarded as being in the market for the good. Failure to account for these differences in the model introduces sample selection bias and may lead to incorrect estimates (Heckman, 1979). On the other hand, one could model the following two equations separately: (i) the decision to buy TMN, and; (ii) willingness to accept the first bid (WTP bid 1), having made the decision to buy a TMN (from i).

However, doing this ignores potential correlations between the error terms of the single equations. For example, a respondent's reasons for not buying a TMN may be correlated with a set of unobservable factors which affect their willingness to accept the first bid. This leads to potential bias in the sampling procedure as the decision on whether to purchase the nets at the first bid or not is not randomly selected. This occurs because the choice to accept the first bid is conditioned on the choice to buy nets. As described by Woolridge (2002), the distribution of the decisions to accept the first bid could be referred to as incidental truncation. To correct for possible sample selection bias, a heckman selection model is specified for the analysis (Heckman, 1979).

In the Heckman selection model, the first part represents the selection model while the second part is the elicitation model. The models were estimated as follows:

1. The selection model specifies those who are willing to buy nets (WTB)

   $\Pr(WTB_i = 1|x) = p(WTB1_i^* > 0|x)$ where,

   Pr denotes the probability and x represents the vector of regressors which are assumed to influence the decision to purchase the valuation good. The model assumes that the error terms are independent and normally distributed.
2. The elicitation model specifies those who are willing to pay for the valuation good at the first bid (Yes to WTP1BID).

   $\Pr(WTP1BID_i = 1|x) = p(WTP1BID_i^* > 0|x)$ where,

   Pr denotes the probability and x represents the vector of regressors which are assumed to influence the decision to purchase TMN at the first bid

(WTP1BID). The model assumes that the error terms are independent and normally distributed.

In the first probit model, the probability that a respondent was willing to buy TMN (WTBNETS) or not, was determined. A selection term, lambda, was saved from this equation and included in the second probit model which estimates the probability that a respondent was willing to pay for TMN at the first bid level (WTP1BID).

To ensure that the estimates for the two equations are unique, exclusion criteria based on the results of the bivariate regression analysis were determined. For this analysis, the variables "number of people living in the house" and "the total number of mosquito measures known" were included in the first probit but excluded in the second probit model. Based on the bivariate regression analysis, both variables influenced the first dependent variable (willingness to buy nets), but not the subsequent dependent variables (willingness to accept the first and second bids).

If selection bias is not established, a two-part model is specified. This allows for conditioning of responses on the choices made to an earlier question. In this case, decision at the first bid (WTP bid 1) is only observed for respondents who choose to buy TMN (WTBNETS). Two-part models treat the two equations as separate and unrelated and models them separately (Frondel & Vance 2012; Belotti et al. 2015). By running the two-part model, the analysis determines both the factors influencing the decision to buy or not buy the TMN and the willingness to pay for the good at the first bid.

Willingness to pay for one TMN at the second bid

All the respondents who answered to the first bid question were presented with a second bid. The bid amounts were increased or decreased, depending on their choices to the first bid (Yes or No) (see figure 7.1). Responses to the second bid are therefore contingent on the response to the first bid (WTPBID1). Failure to observe this dependence in the analysis leads to an incorrect estimation of the WTP bid. For these analyses, two-part models discussed above were estimated. The estimations were done separately for the two responses at this bid path. The two separate sets of analysis were estimated as follows:

(A) Willingness to pay (WTP) at the second higher bid;

1. The first probit model specifies those who are willing to pay for a TMN at the first bid (Yes to WTP1BID).

   $\Pr(WTP1BID_i = 1|x) = p(WTP1BID_i^* > 0|x)$ where,

2. The second probit model (Respondents WTP at the second higher bid (WTPBIDy):

   $\Pr(WTPBIDy_i = 1|x) = p(WTPBIDy_i^* > 0|x)$ where,

Pr denotes the probability and x represents the vector of regressors which are assumed to influence the decision to say yes to the second higher bid (WTPBIDy). The model assumes that the error terms are independent and normally distributed.

(B) Willingness to pay (WTP) at the second lower bid

1. The first probit model specifies those who are willing to pay for a TMN at the first bid (Yes to WTP1BID).

   $\Pr(WTP1BID_i = 1|x) = p(WTP1BID_i^* > 0|x)$ where,

2. The second probit model (Respondents WTP at the second lower bid (WTPBIDn):

   $\Pr(WTPBIDn_i = 1|x) = p(WTPBIDn_i^* > 0|x)$ where

Pr denotes the probability and x represents the vector of regressors which are assumed to influence the decision to say yes to the second lower bid (WTPBIDn). The model assumes that the error terms are independent and normally distributed.

8.2.5 Model diagnostics

The linktest (Cameron & Trivedi 2009) was used to examine specification errors in the models. This test works by creating a variable of prediction and a second one, of the squared prediction and fitting the specified model with the two variables. When a regression model is well specified, the coefficient of the variable of the squared significance should not be statistically significant[35]. Further, the Hosmer Lemeshow test was used to check for the goodness of fit of the models (Archer & Lemeshow 2006; Hosmer & Lemeshow 2013)[36].

Finally, collinearity tests were conducted for the independent variables to check whether they were within tolerable ranges (Chatterjee et al. 2000; Gujarati 2003). Specifically, the variable inflated factor (VIF) and tolerance were used. As suggested in the literature, variables with a VIF greater than 10 and tolerance lower than 0.1 were investigated further and decisions taken based on the influence of specific variables on the model.

8.3 Results

In this section, the results of the analysis are presented. This begins with descriptive statistics and is followed by the results of the regression analyses.

8.3.1 Descriptive statistics

A total of 1,200 respondents were approached for the survey in equal proportions of 300 for each of the four intervention groups: (i) the treated mosquito net (TMN), (ii) in-house residual spraying (IRS), (iii) the active case detection and treatment (ACDT) group and, outside the trial area (OTA). The survey response rate was 100% for the general survey questions (excluding responses to the WTP questions which are covered in a later section). In the following sections, the characteristics of the entire study sample are presented by the independent variables clusters outlines in section 8.2.2. Comparisons of the intervention groups are presented within the discussions. Additional analysis by intervention group is presented in Appendix 23.

[35] The linktest is based on the idea that if a model is properly specified no additional independent variables should be significant above chance. The link test looks for a specific type of specification error called a link error wherein a dependent variable needs to be transformed (linked) to accurately relate to independent variable. The link test adds the squared independent variable to the model and tests for significance versus the non-squared model. A model without a link error will have a nonsignificant t-test versus the unsquared version.

[36] The Hosmer Lemeshow test measures how well a model is specified by grouping cases together according to their predicted values from the logistic regression model. These predicted values are arranged from the lowest to the highest and separated into several groups of approximately equal size. For each group, the observed number of events and non-events is calculated, and the expected number of events (the sum of predicted probabilities for all the individuals in the group) and non-events (the group size less the expected number of events) too. The observed counts and expected counts are then compared using Pearson's chi-square. Low p-values (significance level, usually set at 0.05) suggest a poor fit of the model and thus it should be rejected while a high p-value suggests a good fit of the model.

1) Respondent and household characteristics

The study setting predominantly comprised of Hindus from the scheduled tribe, and who live in Kaccha type of houses. There were no differences across the intervention groups by sex and wealth indicators. However, there were significant differences by attributes such as whether the respondent was the main earner or not, the household religion and caste, and, the type of house the household lived in. Across the intervention groups, the respondents were primarily the main earners. In addition, the majority of the respondents were from the scheduled tribe and practiced the Hindu faith. Further, most of the study respondents lived in kaccha type of houses. Table 8.1 summarises the key respondent and household characteristics.

2) Socio-economic characteristics

In addition to household characteristics, socio-economic attributes relating to both the respondent and the main earner were collected. Table 8.2 summarises the education and occupation characteristics. Significant differences were noted across the intervention groups by the respondent and main earner's education level (Appendix 23).

Table 8-1: Respondent and household characteristics

Attribute	**Summary**	Attribute	Summary
Study respondents		*Household religion*	
Male	95.5%	Hindu	89.58%
Main earners	88.92%	Christians	9.57%
Decision Maker	43.06%	Muslims	0.67%
Household composition		Parsees	0.08%
Household size	Av. 5 (range 1-23)	*Household Caste*	
#children <5 years	Av. 1 (range 0-5)	Scheduled Tribe	74.50%
#Adult Males	Av. 2 (range 0-8)	Backward Caste	17%
#Adult Females	Av. 2 (range 0-7)	Scheduled Caste	3%

Table 8-2: Education and occupation attributes

Variable	Category	Respondent (%)	Main Earner (%)
Education	Illiterate	573 (47.75)	553 (46.08)
	Primary	393 (32.75)	402 (33.50)
	Secondary	206 (17.17)	216 (18.00)
	Graduate	28 (2.33)	29 (2.42)
	Total	**1,200 (100.00)**	**1,200 (100.00)**
Occupation	Agriculture	281 (23.42)	275 (22.92)
	Animal husbandry	81 (6.75)	73 (6.08)
	Labour work	545 (45.42)	536 (44.67)
	Service	163 (13.58)	216 (18.00)
	Business	78 (6.50)	83 (6.92)
	Others	52 (4.33)	17 (1.42)
	Total	**1,200 (100.00)**	**1,200 (100.00)**

The average annual income for the study respondents was Rs. 30,766 (SD: 54,218; range: 0-807,200) obtained primarily from wages (60%) and agricultural activities (44%). The majority of the study respondents were in the 50th income percentile, earning an annual income of between Rs.3, 000 and Rs.16, 000. For the majority of the households, cash income is most available in the months of October (25%), November (20%) and December (12%).

Annual household expenses averaged Rs. 31,515 (range 2,565 – Rs.347, 210). For most households, the highest expenditure items were food and drink, and agricultural expenses (mean annual expenses: Rs. 12,211 and Rs. 4,739). The annual mean expense for health care was Rs. 746.

3) Mosquito and Malaria variables

Knowledge and attitudes towards mosquitoes and prevention methods

More than three quarters (76.33%) of the respondents indicated that mosquitoes were a major nuisance for them and the differences across the intervention groups were significant (Appendix 23). Study respondents were aware of an average of 4 mosquito prevention methods (range 1-8).

The average number of mosquito prevention methods used by the respondents was two (range: 0-7). Table 8.3 summarises the proportion of respondents who know and/or use different malaria prevention methods other than the TMN. The average monthly and total expense for the different mosquito prevention methods is also summarised for the mosquito and non-mosquito season. The use of ITNs and the related expenses are discussed in a later section.

Table 8-3: Knowledge, use and cost of malaria prevention methods

Malaria prevention method	Knowledge n (%)	Use n (%)	Mean monthly expense [Rs.] (SD)		
			During mosquito season	Outside mosquito season	Monthly total
Mosquito coil	235 (19.58)	36 (3)	40.11 (43.11)	12.08 (22.16)	52.19 (62.06)
Mosquito mats	88 (7.33)	15 (1.25)	44.93 (26.90)	22.4 (20.58)	67.33 (44.05)
Smoke*	1,065 (88.75)	984 (82)	-	-	-
Odomos	30 (2.50)	5 (0.42)	44 (32.09)	18 (20.49)	62 (50.69)
Use of oil	52 (4.33)	22 (1.83)	-	-	-
Use of sheets*	928 (77.33)	929 (77.42)	-	-	-
Use of fan*	545 (45.42)	432 (36)	-	-	-
Other methods	13 (1.08)	5 (0.42)	93 (67.60)	33 (42.66)	126 106.96)

*Expenses related to the use of these methods were not documented in the study

Knowledge of and exposure to Malaria

Nearly all the respondents (96.92%) knew of at least one disease caused by mosquitoes. However, only approximately one-fifth (15%) could name the diseases. Other characteristics related to knowledge and exposure to malaria are summarised in table 8.4. There were no significant differences in these attributes across the intervention groups (Appendix 23).

Further, during a malaria episode for one of the family members, two-fifths (40.91%) of the households used cash income to pay for treatment while one third used their savings, sold animals or borrowed funds. Only two households received reimbursement for expenses incurred during the treatment of a malaria episode from their employer, state, government or other organization while one household paid for part of the treatment in kind. The mean amount of money spent for each funding source is summarised in table 8.5.

Table 8-4: Knowledge of and exposure to malaria

Characteristics	Summary
Knowledge of diseases caused by malaria	
Any disease (not specified)	96.92%
Malaria	6.10%
Fever	3.53%
Boils	5.07%
Months when malarial fever most prevalent	
July	15.25%
August	42.08%
September	17%
Malaria incidence in household	
Av. # of household members affected last year	2 (range: 0-8)
Malaria diagnosis	
Doctors	56.16%
Self-diagnosis	30.51%
Sources of treatment for malaria	
Private clinics	57.2%
Private hospitals	17.94%
Public health centres	14.52%
Malaria incidence expenditure and losses	
Av. Expenditure per malaria episode	Rs.145 (SD 337.69)
Av. Days lost due to malaria (patient)	5
Av. Days lost due to malaria (family members)	2
Av. Household income lost due to malaria	Rs. 134

Table 8-5: Mean amounts used to treat malaria by funding source

Funding source	Mean amount Rs. (SD)
Cash income	66 (211.97)
Savings and investments	14.83 (68.56)
Sold goods or animals	6 (67.44)
Borrowing	54.42 (195.75)

Treated malaria net variables

Knowledge, ownership, use and cost of mosquito nets

The majority of study respondents (92.50%) mentioned mosquito nets as a prevention method. This was indicated as the first preferred method by most respondents in the TMN group, compared to the other groups and the difference was statistically significant. For the second preferred method, the use of smoke was mentioned by nearly half (n=139) of the respondents in the TMN group and this difference was statistically significant across the intervention groups.

Table 8.6 summarises the households net use patterns and cost for up to five nets owned by the household.

Table 8-6: Household mosquito net use patterns

TMN Variable	Summary
Owned mosquito nets/household (Av. 2)	41.58%
No. of TMN owned right for family	57.72%
Av. Project nets/household in project sites	1.6
Av. Nets Purchased through the market/household	0.7
Av. Price of TMN	Rs. 123 (SD. 51.70)
Av. # of TMNs required/household	2.4 (SD 1.2)

Nearly all (n=297) of the households in the TMN group owned and used nets when compared to only one-sixth of the ACT and IRS and one-fifth of the OTA. This difference was statistically significant. Further, households in the TMN group owned an average of 3 TMNs, with nearly all distributed through the project while the other groups owned an average of 1 TMN or less. None of the nets owned by the OTA group were distributed through the project and this is a confirmation that this sample was not contaminated and therefore a good comparison.

Approximately one-third (29.58%) of the study households had been sprayed completely in the year of the study, compared to more than three-quarters. Up to six times more households were not sprayed completely in the study year (62.92%) compared to the previous year (11.08%). Spraying of households was not different across the intervention groups.

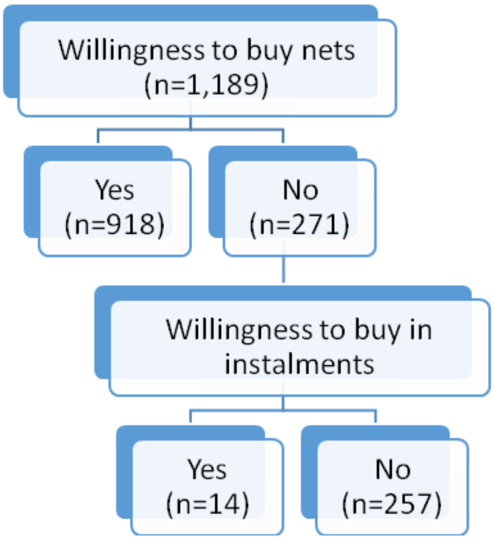
WTP Valuation results

From the total sample of 1,200, four respondents who indicated DK or N/A for all the WTP questions were dropped from the analysis. In addition, for a further 7 respondents who provided their WTP values, their responses were inconsistent (e.g. expressing a lower amount for the total WTP for all the household nets than their WTP for one TMN). These suggested inconsistencies in their preferences probably occasioned by a lack of understanding of the valuation process. They were therefore dropped from the analysis. Therefore, the analysis is based on a reduced sample of 1,189 respondents which represents 99% of the sample. The eleven respondents dropped from the analysis are not expected to affect the internal validity of the estimates (Morton et al. 2005; Fincham 2008; Bhattacherjee 2012). The descriptive results of the valuation process are presented in the next section.

1) Willingness to buy TMNs

More than three-quarters (77.21%) of the respondents were willing to buy treated mosquito nets. When offered the opportunity to pay for the nets in instalments, fourteen more were willing to buy the treated mosquito nets. Overall, 78.39% (n=932) of the respondents were willing to buy treated mosquito nets in either cash or through instalments (figure 8.1). In comparisons across the intervention groups, more than two-thirds (over 80%) of the respondents in the ACDT, IRS and OTA groups were willing to buy TMNs compared to slightly more than half of the respondents in the TMN group (53%) [Appendix 23]. All the respondents who were willing to buy nets either in cash or instalments were taken through the valuation process. The bid path analysis is presented in the next section.

Figure 8-1: Willingness to buy nets in cash and instalments

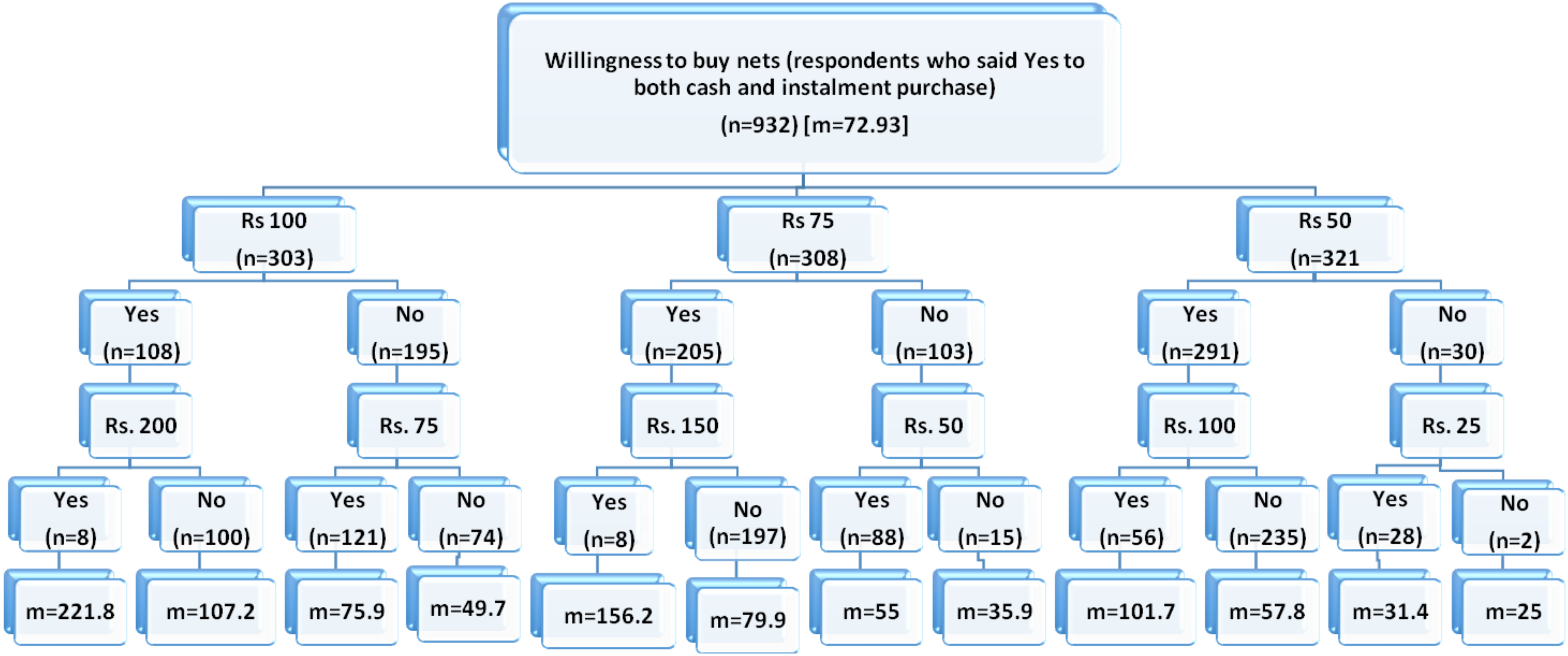

Willingness to buy nets (n=1,189)
Yes (n=918)
No (n=271)
Willingness to buy in instalments
Yes (n=14)
No (n=257)

2) Willingness to pay (WTP) for TMN (Bid path analysis)

Nearly two-thirds (64.81%) of the respondents accepted the first bid offered to them (*Rs.100*: 35.6%; *Rs.75*: 66.55%; *Rs.50*: 90.65%). Of these, approximately more than half of respondents in the other groups were willing to pay for the nets at the first bid amounts, compared to approximately two-fifths (37.92%) of the TMN group and this difference was statistically significant. Thirty five percent of the respondents (n=328) rejected the first bid.

Among the respondents who accepted the first bid, only 7.73% accepted the second higher bid, with the majority (92.27%) rejecting it. Only 2.68% of respondents in the TMN group accepted this bid. However, among those who rejected the first bid, the majority (90.24%) accepted a lower second bid with only 91 (9.76%) rejecting it (Appendix 23). There were no differences in this across the intervention groups. Figure 8.2 illustrates the responses at each stage of the bid and the mean WTP from the final open-ended question.

Figure 8-2: Bid path responses and maximum WTP for 1 treated mosquito net



3) Cumulative bid path responses

Using the bid path responses in figure 8.2, the cumulative bid-acceptance responses are determined by aggregating the responses from the highest to the lowest bid amounts. In this, respondents who were offered the Rs.100 bid, accepted the lower Rs.25 bid; those who accepted the Rs.150 bid accepted the lower Rs.50 bid; and those who accepted the Rs.200 bid accepted the lower Rs.75 bid. Cumulatively, slightly more than half the respondents (52.6%) were willing to pay at least Rs.75 for one TMN.

As illustrated in table 8.7, the probability of accepting a higher bid (from among the low bids) was higher at the higher starting bids.

Table 8-7: Cumulative bid response rates

Response amount (Rs.)	**Initial bid amount (Rs.)**		
	50	75	100
≥25	0.994		
≥50	0.906	0.951	
≥75	0.175	0.662	0.755

8.3.2 Model estimation results

For ease of discussion, I focus on the outputs from the reduced models as these present a better specification and fit. Model estimation results are presented by dependent variable below while the detailed results are discussed in a later section. The results from the analysis of the combined bids are presented first. The output from the analysis of separate bids is used to further support the discussions in relevant sections.

8.3.2.1 Willingness to buy nets

For this dependent variable, analysis was conducted using only the combined model. The analysis included both the respondents who were in the market for TMNs and those who were not. Respondents were subjected to the different bid paths only after this stage. The reduced model estimation results for this dependent variable are presented in table 8-8 while the univariate and base model estimation output is presented in Appendix 24.

Among this sample, willingness to buy nets was increased by nearly half (47.1%) for respondents who belonged to the scheduled tribe. Respondents who did not consider mosquitoes to be a major nuisance were least likely to be WTP for TMN

(88.5%) with only approximately half (50.6%) of those who were classified as being in the “other” occupation expressing WTB the nets compared to those in the agriculture sector.

Table 8-8: Model estimation results _Willingness to buy nets

Variable (Base)	**Categories**	**Coefficient[#] (Robust standard error)**
Interview village (Treated mosquito nets)	Active Case Detection	1.144***(0.211)
	In-house spray village	1.247***(0.206)
	Outside trial area	1.054***(0.209)
No. of people living in the house		0.0477*(0.0244)
Sex of the main earner (Male)		-0.299*(0.161)
Respondent’s Main Occupation (Agriculture)	Animal Husbandry	
	Business	
	Labour Work	-0.296***(0.114)
	Others	-0.506**(0.226)
	Service	
Caste household belongs to (Scheduled caste)	Other backward caste	
	Scheduled tribe	0.471***(0.160)
Type of house (Kaccha)	Pucca	
	Semi pucca	-0.460*** (0.109)
Whether respondent considers mosquitoes to be a nuisance (Major nuisance)	Minor nuisance	
	No Nuisance	-0.885***(0.267)
No. of nets purchased from market		-0.148**(0.0609)
Expenditure incurred on malaria treatment (log)		0.0896***(0.0307)
Constant		0.00521(0.322)
Pseudo R^2		0.1761
Linktest		0.85601[a]
Goodness of fit		4.41[b]
Observations		1,189

# The estimated parameters and asterisks show significance level of 1% (***), 5% (**) and 10% (*)

[a] p=0.398 [b] p=0.8182

8.3.2.2 Willingness to pay at the first bid

This analysis included both the respondents who were in the market for TMNs and those who were not (n=1,189). This was done to test whether there was any selection bias among respondents who chose to participate in the valuation process compared to those who did not. As figure 8.2 shows, respondents were allocated to one of three starting bids (Rs.50, Rs.75, and Rs.100). This was a binary choice question with two possible outcomes for each bid: Yes (WTPBID1=1) if this first bid is accepted and (2) No (WTPBID1=0) if the first bid is not accepted as discussed in chapter 7.

In the next section, the results of the combined bids are presented. The same analysis is conducted for the different bid paths and this is presented in a later section as additional analyses [tables 8.16 - 8.17 (bid Rs.50), tables 8.18 – tables 8.19 (bid Rs.75) and tables 8.20 - 8.21 (bid Rs.100)].

The correlation coefficients between the error terms of the two probit models for dependent variable 2 (willingness to pay for TMN at the first bid, given the decision to buy TMN) were not statistically significant (0.8632). Selection bias was therefore not detected, and the two equations were modelled separately (Appendix 27). The reduced model estimation results are presented in table 8.9 while the univariate and base model estimation outputs are presented in Appendix 26.

Table 8-9 Model estimation results _Willingness to pay for nets bid 1

Variable (Base)	**Categories**	**Coefficient[#] (Robust standard error)**
No. of people living in the house		0.0556**(0.0220)
Whether respondent main earner or not (Yes)		0.269*(0.142)
Main earner's education (Graduation)	Illiterate	-0.329***(0.123)
	Primary	-0.249**(0.110)
	Secondary	
Respondent's Main Occupation (Agriculture)	Animal Husbandry	-0.322*(0.192)
	Business	-0.532***(0.169)
	Labour Work	-0.515***(0.117)
	Others	
	Service	-0.350**(0.170)
Caste household belongs to (Scheduled caste)	Other backward caste	
	Scheduled tribe	0.236*(0.125)
Whether respondent considers mosquitoes to be a nuisance (Major nuisance)	Minor nuisance	
	No Nuisance	-0.534**(0.267)
Total no. of mosquito measures known		-0.0711*(0.0420)
Mosquito net among preferred methods (Yes))		0.186*(0.108)
Total no. of prevention methods used		0.164***(0.0612)
If any disease is caused by mosquito bites (No)		0.622**(0.280)
Constant		-0.586*(0.329)
Pseudo R^2		0.1536
Linktest		-0.0784[a]
Goodness of fit		8.73[b]
Observations		932

# The estimated parameters and asterisks show significance level of 1% (***), 5% (**) and 10% (*)
[a] p=0.736 [b] p=0.3665

Willingness to pay at the first bid amount was significantly increased for larger households. The WTP at this bid was also increased by more than one-quarter when the study respondent was also the main earner in the household (26.9%) and by more than one-fifth (23.6%) for households which belonged to the scheduled tribe (compared to those from the scheduled caste). The WTP at this bid was highest (62%) among respondents who knew of a disease caused by mosquito bites.

On the contrary, respondents were least likely to be WTP at this first bid if they were engaged in business (53%) and labour work (51.5%), when compared to those engaged in agriculture as an occupation.

Willingness to pay for TMN at the second bid

From the initial sample of 1,189, more than one-fifth (257) respondents were not willing to buy TMNs either in cash or instalments and they were therefore excluded from this analysis, leaving 932 respondents for the analysis of these dependent variables. As all the respondents were included in the analysis, selection bias was not considered to be a problem. However, the response to this question was dependent on responses to the previous question, as illustrated in figure 8.2 and the results of the two-part models estimated are presented in the next section.

8.3.2.3 Willingness to pay at the second higher bid

The reduced model estimation results of the two-part model for the dependent variable (willingness to pay for nets at the second higher bid) are presented in table 8.10. The univariate and base model estimation results are presented in Appendix 28.

Willingness to pay for nets at the second higher bid was increased by approximately one-third (31%) for households which lived in a pucca structure (compared to Kaccha). This was also positively and significantly influenced by the total number of prevention methods used (19%) and the total number of nets purchased in the market (31.2%). Respondents who engaged in labour work (compared to agriculture) were less likely to be willing to accept this second higher bid (42.2%) with similar rates for households which belonged to the other backward caste (41.8%) when compared to those from the scheduled caste. Purchasing nets from the market also negatively affected respondents' WTP for the TMN at this second higher bid (12%).

Table 8-10: Model estimation results _Willingness to pay for nets (second higher bid)

Variable (Base)	**Categories**	**Coefficient[#] (Robust standard error)**
Respondent's Main Occupation (Agriculture)	Animal Husbandry	
	Business	
	Labour Work	-0.422***(0.124)
	Others	
	Service	
Caste household belongs to (Scheduled caste)	Other backward caste	-0.418***(0.154)
	Scheduled tribe	
Type of house (Kaccha)	Pucca	0.316*(0.186)
	Semi pucca	
Total no. of prevention methods used		0.190**(0.0839)
Whether household owns nets: Yes		
No. of nets owned		-0.120**(0.0590)
No. of nets purchased from market		0.312***(0.0974)
Constant		-1.770***(0.212)
Pseudo R^2		0.1028
Linktest		-0.1583[a]
Goodness of fit		8.84[b]
Observations		932

[#] The estimated parameters and asterisks show significance level of 1% (***), 5% (**) and 10% (*)
[a]=0.415 [b] p=0.2641

8.3.2.4 Willingness to pay at the second lower bid

The reduced model estimation results of the two-part model for the dependent variable (willingness to buy nets at the second lower bid) are presented in table 8.11. The univariate and base model estimation results are presented in Appendix 29.

Willing to pay for nets at this lower bid was reduced by close to two-thirds (58.2%) if the main earner was illiterate compared to the households where the main earner was a graduate and by nearly three-quarters if the main earner's occupation was labour work. However, WTP for nets at this bid significantly increased among respondents from the scheduled tribe (96.4%) and the other backward caste (56.5%). Knowledge of diseases caused by mosquito bites also increased willingness to pay for nets at this bid by more than three-quarters (76.4%).

Table 8-11: Model estimation results _Willingness to pay for nets (second lower bid)

Variable (Base)	**Categories**	**Coefficient[#] (Robust standard error)**
Respondent's Main Occupation (Agriculture)	Animal Husbandry	
	Business	-0.486*(0.248)
	Labour Work	-0.691***(0.161)
	Others	-0.643**(0.312)
	Service	-0.451**(0.193)
Caste household belongs to (Scheduled caste)	Other backward caste	0.515**(0.255)
	Scheduled tribe	0.820***(0.254)
Whether household owns nets: Yes		0.322**(0.144)
If any disease is caused by mosquito bites (No)		0.743***(0.218)
Constant		0.245(0.338)
Pseudo R^2		0.0657
Linktest		0.4401[a]
Goodness of fit		2.72[b]
Observations		932

[#] The estimated parameters and asterisks show significance level of 1% (***), 5% (**) and 10% (*)
[a]=0.162 [b] p=0.8435

The decision to pay for TMN at the second lower bid was positively and significantly influenced by the household castes (51.5% for the other backward caste and 82% for the scheduled tribe). Knowledge of a disease caused by mosquitoes also increased the WTP for TMN by approximately three-quarters (74.3%). All the occupation categories negatively and significantly influenced WTP for one TMN.

In the next section, the results are presented by the four independent variable clusters. Comparisons are made across the four models (dependent variables/ bid levels) to determine similarities or divergence among the predictors of WTP. While significant across all the categories, the interview village does not fit into the below clusters and is therefore not presented in the summaries.

1. Respondent and household characteristics (gender, religion, caste, type of house, household size and number of children in the household, main earner in the household);
2. Socio-economic characteristics (education, employment and income);
3. Malaria variables (knowledge, exposure and experience with the disease, knowledge of and use of prevention methods);
4. Treated malaria net variables (current ownership and source of current net).

8.3.2.5 Respondent and household characteristics

Table 8.12 summaries all the significant respondent and household variables for each of the dependent variables. Evidently, different respondent and household factors influence the willingness to buy or pay for nets at the different bid levels. Among the variables, only the household caste remains a significant factor across the different bid levels. This is also in the hypothesised directions. A household from the scheduled tribe, was 47% more likely to buy a net when compared to those from the scheduled caste. From these, only slightly more than one-fifth (23.6%) accepted the first bid. Approximately two-thirds (41.8%) of households from the other backward caste would not be willing to accept the second higher bid while slightly more than half of these (51.5%) would be willing to accept the second lower bid. More than four fifths (82%) of respondents from the scheduled tribe were willing to accept the second lower bid.

8.3.2.6 Socioeconomic characteristics

A summary of the significant socioeconomic characteristics across the different variables are presented in table 8.13. Only the respondent's occupation status predicted WTP for all the dependent variables. The direction of effect was also consistent with the results aligning with the apriori expectations. The education variable was only significant for the first bid but not the rest. Contrary to apriori expectations, income was not a predictor of WTP in this combined sample.

8.3.2.7 Net variables

None of the net variables predicted WTP for nets across the different valuation levels (Table 8.14). Household net ownership only influenced ownership for the lowest bid while the number of nets owned per household affected preferences at the second higher bid only. The number of nets previously purchased from the market negatively influenced the decision to purchase nets and whether respondents were WTP at the second higher level.

8.3.2.8 Malaria knowledge variables

As with the net variables, among the significant malaria knowledge variables, none influenced the decision to buy or pay for nets across the four valuation levels (Table 8.15). However, all the significant variables influenced WTP in the predicted directions. Respondent's perception of mosquitoes negatively influenced the decision to buy and purchase nets at the first bid but not the rest. On the contrary, respondents who knew of diseases caused by mosquitoes were more likely to accept the first (62.2%) and second lowest bids (74.3%).

These results are similar when the bid paths are analysed separately (Tables 8.6 – 8.21). Along the Rs.50 bid path, the respondents' occupation and whether the respondent considered mosquitoes to be a nuisance remained significant for the first and second higher bids, but not the second lower bid. Along the Rs.75 bid path, the household caste and total number of prevention methods used by the household influenced the decision to WTP for the first and second higher bids. Factors that remained significant along the Rs.100 bid path were the respondents' occupation, net ownership in the household and knowledge about diseases caused by mosquito bites. The magnitude and direction of effect was similar for the first and second lower bids along this bid path.

The estimation results from both the combined and separate bid paths further illustrate the argument presented in this dissertation. Firstly, predictors of WTP are different across valuation levels where multiple techniques are used. Secondly, predictors also differ by bid paths. Given this, summary WTP statistics based on multiple elicitation techniques, and/or multiple bid levels and the WTP predictions thereof are expected to differ. Conclusions on criterion validity assessments are therefore largely driven by the choice of summary statistics presented for this comparison. As discussed in earlier chapters in the dissertation (chapters 4-6), the choice of summary statistic presented for criterion validity comparisons is often not discussed by authors. Where multiple elicitation techniques are employed, the justification for the type of analysis (combined or by different bid paths or valuation levels) is not presented either. This would aid in the interpretation of criterion validity assessments conducted.

Table 8-12: Summary of significant respondent and household characteristics

Variable (Reference Category)	WTB Coef#. (s.e)	WTP (Bid 1) Coef#. (s.e)	WTP (Bid 2 Higher) Coef#. (s.e)	WTP (Bid 2 Lower) Coef#. (s.e)
No. of people living in the house	0.0477*(0.0244)	0.0556**(0.0220)	-	-
No. of children below the age of 6 years	-	-	-	-
Whether respondent main earner or not (Yes)	-	0.269*(0.142)	-	-
Sex of the main earner (Male)	-0.299*(0.161)	-	-	-
Caste household belongs to (Scheduled caste)				
Other backward caste		-	-0.418***(0.154)	0.515**(0.255)
Scheduled tribe	0.471***(0.160)	0.236*(0.125)		0.820***(0.254)
Type of house (Kaccha)				
Pucca		-	0.316*(0.186)	-
Semi pucca	-0.460*** (0.109)	-	-	-

# The estimated parameters and asterisks show significance level of 1% (***), 5% (**) and 10% (*)

Table 8-13: Summary of significant socioeconomic characteristics

Variable (Reference Categories)	**WTB** **Coef$^{\#}$. (s.e)**	**WTP (Bid 1)** **Coef$^{\#}$. (s.e)**	**WTP (Bid 2 Higher)** **Coef$^{\#}$. (s.e)**	**WTP (Bid 2 Lower)** **Coef$^{\#}$. (s.e)**
Main earner's education (Graduation)				
Illiterate	-	-0.329***(0.123)	-	-
Primary	-	-0.249**(0.110)	-	-
Secondary	-	-	-	-
Respondent's Main Occupation (Agriculture)				
Animal Husbandry	-	-0.322*(0.192)	-	-
Business	-	-0.532***(0.169)	-	-0.486*(0.248)
Labour Work	-0.296***(0.114)	-0.515***(0.117)	-0.422***(0.124)	-0.691***(0.161)
Others	-0.506**(0.226)	-	-	-0.643**(0.312)
Service	-	-0.350**(0.170)	-	-0.451**(0.193)
Total household disposable income (log)	-	-	-	-

[#] The estimated parameters and asterisks show significance level of 1% (***), 5% (**) and 10% (*)

Table 8-14: Summary of significant net variables

Variable (Reference Categories)	**WTB** **Coef$^{\#}$. (s.e)**	**WTP (Bid 1)** **Coef$^{\#}$. (s.e)**	**WTP (Bid 2 Higher)** **Coef$^{\#}$. (s.e)**	**WTP (Bid 2 Lower)** **Coef$^{\#}$. (s.e)**
Whether household owns nets: Yes	-	-	-	0.322**(0.144)
No. of nets owned	-	-	-0.120**(0.0590)	-
No. of nets purchased from market	-0.148**(0.0609)	-	0.312***(0.0974)	-

[#] The estimated parameters and asterisks show significance level of 1% (***), 5% (**) and 10% (*)

Table 8-15: Summary of significant malaria knowledge variables

Variable (Reference Categories)	**WTB** **Coef$^{\#}$. (s.e)**	**WTP (Bid 1)** **Coef$^{\#}$. (s.e)**	**WTP (Bid 2 Higher)** **Coef$^{\#}$. (s.e)**	**WTP (Bid 2 Lower)** **Coef$^{\#}$. (s.e)**
Whether respondent considers mosquitoes to be a nuisance (Major nuisance)				
Minor nuisance	-	-	-	-
No Nuisance	-0.885***(0.267)	-0.534**(0.267)	-	-
If any disease is caused by mosquito bites (No)	-	0.622**(0.280)	-	0.743***(0.218)
No. of family members suffering from malaria last month	-	-	-	-
Expenditure incurred on malaria treatment (log)	0.0896***(0.0307)	-	-	-
Total no. of mosquito measures known	-	-0.0711*(0.0420)	-	-
Mosquito net among preferred methods (Yes))	-	0.186*(0.108)	-	-
Total no. of prevention methods used	-	0.164***(0.0612)	0.190**(0.0839)	-

[#] The estimated parameters and asterisks show significance level of 1% (***), 5% (**) and 10% (*)

Table 8-16: Rs.50 bid path model estimation results _ Willingness to pay for nets bid 1

Variable (Base)	**Categories**	**Coefficient[#] (Robust standard error)**
No. of people living in the house		0.137***(0.0527)
Main earner's education (Graduation)	Illiterate	
	Primary	-0.373*(0.225)
	Secondary	
Respondent's Main Occupation (Agriculture)	Animal Husbandry	
	Business	-1.015*(0.583)
	Labour Work	-1.252***(0.417)
	Others	
	Service	-0.914*(0.508)
Total household disposable income (log)		0.121***(0.0329)
Whether respondent considers mosquitoes to be a nuisance (Major nuisance)	Minor nuisance	-0.493*(0.289)
	No Nuisance	-1.456***(0.383)
Total no. of mosquito measures known		0.2876
Mosquito net among preferred methods (Yes))		0.471*(0.279)
Constant		0.461(1.849)
Pseudo R^2		1.195**(0.484)
Linktest		-0.0924[a]
Goodness of fit		4.40[b]
Observations		321

[#]The estimated parameters and asterisks show significance level of 1% (***), 5% (**) and 10% (*)
[a] p=0.635 [b] p=0. 0.8197

Table 8-17: Rs.50 bid path model estimation results _ Willingness to pay for nets (second higher bid)

Variable (Base)	**Categories**	**Coefficient[#] (Robust standard error)**
Respondent's Main Occupation (Agriculture)	Animal Husbandry	
	Business	0.734**(0.310)
	Labour Work	-0.394**(0.183)
	Others	
	Service	
Caste household belongs to (Scheduled caste)	Other backward caste	-0.589**(0.233)
	Scheduled tribe	
Whether respondent considers mosquitoes to be a nuisance (Major nuisance)	Minor nuisance	0.490*(0.250)
	No Nuisance	
No. of nets owned		-0.209***(0.0789)
No. of nets purchased from market		0.629***(0.143)
Constant		-0.934***(0.157)
Pseudo R^2		0.1615
Linktest		-0.124[a]
Goodness of fit		4.56[b]
Observations		321

[#]The estimated parameters and asterisks show significance level of 1% (***), 5% (**) and 10% (*)

[a] p=0.463 [b] p=0.6007

Table 8-18: Rs.75 bid path model estimation results _ Willingness to pay for nets bid 1

Variable (Base)	**Categories**	**Coefficient[#] (Robust standard error)**
Interview village (Treated mosquito nets)	Active Case Detection	
	In-house spray village	
	Outside trial area	-0.331**(0.156)
Main earner's education (Graduation)	Illiterate	-4.949***(0.341)
	Primary	-4.934***(0.318)
	Secondary	-4.752***(0.312
Respondent's Main Occupation (Agriculture)	Animal Husbandry	
	Business	
	Labour Work	-0.309*(0.169)
	Others	
	Service	
Religion of the household (Christian)	Hindu	-0.573**(0.237)
	Muslim	
	Parsee	
Caste household belongs to (Scheduled caste)	Other backward caste	-0.660**(0.264)
	Scheduled tribe	
Type of house (Kaccha)	Pucca	
	Semi pucca	
Whether respondent considers mosquitoes to be a nuisance (Major nuisance)	Minor nuisance	0.680***(0.237)
	No Nuisance	
Total no. of prevention methods used		0.274***(0.101)
Constant		5.460***(0.446)
Pseudo R^2		0.1014
Linktest		0.1032[a]
Goodness of fit		2.64
Observations		308

[#]The estimated parameters and asterisks show significance level of 1% (***), 5% (**) and 10% (*)

[a] p=0.710 [b] p=0.9551

Table 8-19: Rs.75 bid path model estimation results _ Willingness to pay for nets (second lower bid)

Variable (Base)	**Categories**	**Coefficient[#] (Robust standard error)**
Main earner’s education (Graduation)	Illiterate	-4.076**(0.761)*
	Primary	-4.372***(0.671)
	Secondary	-3.150***(0.630)
Caste household belongs to (Scheduled caste)	Other backward caste	
	Scheduled tribe	
Type of house (Kaccha)	Pucca	-0.985*(0.484)
	Semi pucca	
Total no. of mosquito measures known		-0.232*(0.138)
Total no. of prevention methods used		0.384*(0.231)
Whether household owns nets: Yes		-1.905***(0.723)
No. of nets owned		0.779**(0.346)
No. of nets purchased from market		1.213**(0.547)
If any disease is caused by mosquito bites (No)		1.439**(0.580
No. of family members suffering from malaria last month		-0.786**(0.335)
Expenditure incurred on malaria treatment (log)		0.200**(0.0719)*
Constant		4.613***(0.819
Pseudo R^2		0.3404
Linktest		0.2360[a]
Goodness of fit		12.03[b]
Observations		235

[#] The estimated parameters and asterisks show significance level of 1% (***), 5% (**) and 10% (*)

[a] p=0.559 [b] p=0.1500

Table 8-20: Rs.100 bid path model estimation results _ Willingness to pay for nets bid 1

Variable (Base)	**Categories**	**Coefficient[#] (Robust standard error)**
Interview village (Treated mosquito nets)	Active Case Detection	
	In-house spray village	
	Outside trial area	0.411**(0.175)
Sex of the main earner (Male)		-1.004**(0.399)
Respondent's Main Occupation (Agriculture)	Animal Husbandry	-0.676**(0.344)
	Business	0.697***(0.203)
	Labour Work	-0.613***(0.179)
	Others	
	Service	
Religion of the household (Christian)	Hindu	0.697***(0.203)
	Muslim	
	Parsee	
Whether household owns nets: Yes		0.534***(0.178)
No. of nets owned		
No. of nets purchased from market		
If any disease is caused by mosquito bites (No)		0.910*(0.538)
Constant		-0.806(0.7012)
Pseudo R^2		0.1159
Linktest		-0.3348[a]
Goodness of fit		3.82[b]
Observations		303

[#]The estimated parameters and asterisks show significance level of 1% (***), 5% (**) and 10% (*)

[a] p=0.211 [b] p=0.8001

Table 8-21: Rs.100 bid path model estimation results _ Willingness to pay for nets (second lower bid)

Variable (Base)	**Categories**	**Coefficient# (Robust standard error)**
No. of people living in the house		-0.0861**(0.0430)
Whether respondent main earner or not (Yes)		0.459*(0.252)
Respondent's Main Occupation (Agriculture)	Animal Husbandry	
	Business	
	Labour Work	-0.572***(0.201)
	Others	
	Service	
Caste household belongs to (Scheduled caste)	Other backward caste	0.645**(0.275)
	Scheduled tribe	0.925***(0.277)
Whether household owns nets: Yes		0.479**(0.199)
If any disease is caused by mosquito bites (No)		0.815*(0.432)
Constant		-0.728(0.589)
Pseudo R^2		0.0911
Linktest		0.0482[a]
Goodness of fit		8.18[b]
Observations		303

#The estimated parameters and asterisks show significance level of 1% (***), 5% (**) and 10% (*)

[a] p=0.865 [b] p=0.419

8.3.3 Predicted WTP values

The predicted probabilities of willingness to pay for TMN were different across the bid levels. Table 8-16 below illustrates this. At the first bid, the probability of buying the TMN was close to the percentage of people who were willing to purchase the nets at the same bid (64.81%). However, the predicted probabilities were different from the hypothetical survey responses for the rest of the bids.

Table 8-22: Predicted probabilities of WTP at different bid levels

Dependent Variable	**Predicted probability**	**Std. Error**	**95% Conf. Interval**
First bid	0.64	0.0051	[0.630, 0.663]
Second higher bid	0.21	0.0623	[0.159, 0.238]
Second lower bid	0.87	0.0041	[0.864, 0.891]

These results further illustrate the different analytical methods that can be used with discrete choice WTP data. In addition, the multiple estimates at the different bid levels are demonstrated. As WTP can be predicted at the different levels, an exploration of criterion validity employing the full range of values is likely to generate different results, with validity confirmed at some of the levels and not others.

In the next section, all the analysis results are discussed and the implications of these on criterion validity highlighted.

8.4 Discussion of results

In this section, the results presented above are discussed considering the apriori evidence on variables expected to influence WTP for TMN. In addition, discussions based on the expected economic and exploratory variables are presented. The final discussion addresses the key concerns summarised in section 8.2.1 which relate to criterion validity conclusions based on multiple estimates of WTP such as the two-stage bidding method presented in this chapter. The findings from both the combined and separate bid path analysis, where relevant, are discussed.

Willingness to buy nets (WTB) is only positive for larger households, those that belong to the scheduled tribe and this increased expenditures on malaria treatment too. However, as expected, respondents who did not consider mosquitoes to be a nuisance were least likely to be willing to buy them (88%) as were those who engaged in labour work compared to agriculture (50.6%) and those who lived in semi pucca structures compared to Kaccha (46%).

At the first bid, respondents are most likely to be willing to pay for TMN if they knew of a disease caused by mosquitoes (62.2%), where the respondent was also the main earner and with increasing household sizes. However, respondents were least likely to be willing to pay for TMN at this bid if they did not consider mosquitoes to be a nuisance, engaged in businesses (compared to those in agriculture) and it the main earner in the household was illiterate.

At the second higher bid, respondents were more likely to be willing to pay for TMN if they lived in a pucca structure (compared to Kaccha). Willingness to pay for nets at this bid also increased with the total number of nets purchased from the market by the household. At this bid level, respondents were least likely to be willing to pay for TMNs if they engaged in labour work or the household belonged to the other backward caste.

Willingness to pay for TMN at the second lower bid was highest among respondents who belonged to the scheduled tribe and those who knew of a disease caused by mosquitoes. Respondents who engaged in labour work were least likely to be willing to pay for TMN, compared to those in agriculture. Overall, compared to agriculture, respondents in all other categories were least likely to be willing to accept the second higher bid.

Evidently, across the different bids, different factors influence WTB and WTP for TMNs. As discussed in earlier sections, respondents who were not willing to buy nets were not taken through the valuation process. It is interesting to note that even these respondents have carefully constructed preferences, which accord to economic theory and some of the exploratory variables. For example, previous incidences of malaria within the household and expenditure incurred in treating the disease positively influenced the decision to buy nets. This finding is similar to previous assessments of WTP for TMN (Onwujekwe et al., 2004). However, this factor did not influence the decision to pay for nets at the different bid levels for the combined bids analysis. Considering the separate bid paths, the expenditure incurred in treating malaria was only positive and significant for the second lower bid along the Rs. 75 bid path. In their study on altruistic WTP for TMN, Onwujekwe and

Chima (2001) established that higher expenses for malaria treatment were positively correlated with WTP. The fact that a value (bid) was not presented in the first valuation question (willingness to buy nets) may have introduced inconsequentiality[37], affecting the response to this question. There is evidence to support the argument that respondents are more likely to respond truthfully to elicitation questions where they believe that their response is consequential (Carson and Groves, 2007). In their empirical research, Vossler et al. (2013) concluded that single dichotomous choice elicitation methods provide sufficient conditions for consequentiality and are thus considered incentive compatible. It is likely that the outcome would have been different if all the respondents were included in the valuation process.

The descriptive results presented in section 8.3 indicated that nearly two-thirds (64.81%) of the study respondents were willing to pay for TMN at the first bid. As evidenced in the variables that influence WTP in the same section, higher household incomes were associated with increased WTP. Household income, also considered an economic variable in this analysis, did not significantly influence WTP in the analyses presented in this chapter. Household expenditure was used as a proxy for income with increasing expenditure taken to imply higher income. The nature of the employment of both the respondent and main earners was also considered as proxies for income. Regular and professional jobs would be expected to generate higher incomes and vice versa.

The results suggest that households where the main earner worked in business or labour work were less likely to be WTP for TMN across the combined bid levels. This is compared to respondents in agriculture. However, in the analysis of the separate bid paths, respondents who engaged in business were more likely to be WTP for TMN for the first bid (Rs.100 bid path) and second higher bid (Rs.50 bid path). A plausible explanation for this finding would be that respondents in this occupation are likely to have higher incomes, which is not seasonal too. Results on the occupation variables highlight the economic status of the households and the effect of this on decision making. The seasonal and uncertain nature of occupations such as labour

[37] Consequentiality is assumed if a respondent believes that their choices or responses to the elicitation question may enter their utility function, in this case, that there would be budgetary implications.

work and service industry would imply lower household disposable incomes. This also points to the fact that cash may not be available at the time of the interview and therefore decisions are made with this in mind. Households where the respondent or main earner worked in the other industry (this included professions such as teaching) were more likely to be willing to pay for TMN at the second higher bid. This result accords with the consumer theory on income and consistency in the decision-making process. The other backward caste is composed of the poorest segment of the population in the study setting. It is therefore not surprising, and in line with economic theory, that respondents from this population group would be willing to pay the least for TMNs.

Further, as evidenced from other studies, females were less likely to be willing to buy nets in the first valuation question. Further, respondents who were main earners were also more likely to be willing to pay for TMN at the first bid. The findings on main earners was the same for the Rs.100 bid path analysis where female respondents were least likely to be WTP for TMN at the first bid with respondents who were also main earners more likely to be WTP for TMN at the second lower bid. These findings are also supported in the literature on WTP for TMN (Onwujekwe & Obinna 2002, Onwujekwe et al, 2004). In the study setting, the decision makers and budget holders within households were male and hence the results further accord with empirical evidence.

Nearly all the exploratory variables included in this analysis conformed to expectations across both the combined and separate bid path analysis. However, as discussed variedly, the results were not consistent across the models. For example, the number of nets owned is supposed to elicit mixed results. Households which already owned TMNs may not need additional nets. However, a positive experience with TMN may also increase their WTP. The results from this variable were different across both the combined and separate bid path analysis. For example, the number of nets owned negatively influenced respondents WTP at the second higher bid along the Rs.50 bid path, and the combined bid analysis. The reverse was true for responses to the second lower bid along the Rs.75 bid path. Both sets of results confirm these hypotheses.

Experience with mosquito prevention methods may increase respondents' knowledge and valuation of the TMN and thus increase WTP. Mixed results were obtained for this variable across the combined and some of the separate bid path analysis. The total number of mosquito prevention methods known negatively influenced respondents' willingness to buy a TMN in the combined analysis. Considering the separate bid path analysis, respondents were more likely to be WTP for a TMN at first bid along the Rs.50 bid path with the reverse holding true for the second bid along the Rs.75 bid path. Where this was identified as a significant predictor, the knowledge of diseases caused by mosquitoes positively influenced the decision to buy or pay for TMN. The apriori expectation was that willingness to buy nets and willingness to pay for nets would be higher among respondents who knew that mosquitoes caused malaria, and vice versa.

Nearly all (92.27%) of the respondents offered the second lower bid were willing to buy nets at this bid. At this bid, which presents the lowest amount offered, WTP for nets was higher if the household belonged to the backward caste or schedule tribe, which are the most deprived sections of this community. This result was true for both the combined bid and the Rs.100 bid path analysis. This finding further accords with the exploratory variables.

Overall, the effect of education on the decision to buy nets or pay for them is mixed. Used as a proxy for income and knowledge, (higher education translates to employment with higher income and a greater understanding of both the elicitation process and the malaria context), the interpretation of the results for the willingness to buy variable are mixed. While the effect of education was not significant in the combined bid analysis, this was identified in the analysis of the separate bids. At the first bid level, compared to graduate level of education, households where the main earner was illiterate, or had primary or secondary level of education were least likely to be WTP for TMN (Rs.50 and Rs.75 bid paths). The same result holds for responses to the second lower bid along the Rs.75 bid path. A plausible explanation for this could be that respondents with graduate level of education could afford to live in more affluent types of houses where the threat of mosquitoes was minimised, hence the decision not to buy nets.

The descriptive analysis presented in section 8.3 highlighted the different acceptance rates across the bid levels, and the filter question. The regression analysis in section 8.4 demonstrates consistency with empirical evidence and theoretical expectations on the factors expected to determine decisions to buy or pay for TMN, at all levels. Respondents did not self-select into the sample as evidenced by the lack of selection bias in the models. However, the predictors of WTP are significantly different, and with minimal overlap across the bid levels and elicitation methods. These have been summarised in tables 8.12 - 8.15.

It is likely that some of the results obtained would have been different, with different specifications of the variables and the estimated models. However, as a secondary dataset was used, analysis was limited to the variables in the dataset. Further, model diagnostics indicated that the regression models were well specified.

It is also likely that different results would have been obtained in a different setting. For instance, if such a study was conducted in a setting where health care is publicly funded, it is possible that the results would have been different. Possibly, protest responses would have been obtained among people opposed to paying for any aspect of health care. However, in the study setting, while health care is provided by the state, out of pocket payment for health care was the norm as at the time of the interviews, as is the case in most low-income settings. As a result, it is likely that the responses reflected the true valuation of the treated mosquito net by the respondents.

Further, one could argue that the findings from this study do not hold currently, given that the study was conducted more than a decade ago. While this may be true for the welfare estimations for the valuation good, the validity of the dataset for this methodological analysis still holds. In particular, the analysis presented in this chapter did not seek to establish the predictors of willingness to buy or willingness to pay for treated mosquito nets as an end to itself. Rather, the purpose of the different analyses presented in this chapter is to highlight the multiple estimates and different predictors of WTB and WTP that can be obtained with the use of multiple WTP elicitation techniques. The impact of such analyses on criterion validity conclusions is discussed in the next section.

8.5 Implications for criterion validity

As discussed in chapters 6, criterion validity comparisons involve comparisons of hypothetical WTP estimates and actual values. Where hypothetical WTP values are elicited using multiple methods, or using formats which involve elicitation at multiple levels, only one summary measure is presented for this comparison. The analysis presented in this chapter has demonstrated the following:

a. The predictors of WTP at different bid levels are indeed different, and the differences are significant;

b. The predicted WTP at the different bid levels are different;

c. Hypothetical WTP can be summarised in multiple ways, arriving at distinctly different summary measures.

Given this, conclusions on criterion validity would be expected to be different, depending on the choice of hypothetical WTP comparator. Therefore, criterion validity assessments based on a single summary estimate from multiple elicitation points are likely to lead to incorrect conclusions on criterion validity.

8.6 Conclusion and chapter summary

Where multiple elicitation techniques are utilised to elicit WTP values, criterion validity assessments for the different estimates would provide more accurate conclusions. However, this is rarely the case and might be considered impractical too. As discussed in previous sections, even where WTP estimations are conducted at multiple points the majority of studies summarise hypothetical WTP into a single estimate, which is then compared with data collected from actual surveys. The results obtained in this chapter highlight the variety in analytical methods, estimates and predictors of WTP; all of which impact criterion validity assessments and conclusions thereof. Mean WTP can also be estimated from discrete choice data. In furthering the discussion on the alternate estimates of WTP, and the effect of these on criterion validity assessments, the analysis of open ended and interval data is considered in the next chapter. Combined, these analyses further highlight the potential flaws with the criterion validity assessments and conclusions thereof. This conclusion is based on the analysis of hypothetical WTP data and the choice of comparator for criterion validity assessments.

Chapter 9 Alternate Estimations of Mean WTP from Open Ended and Interval data

9.1 Introduction

In chapter 8, the analysis of hypothetical WTP data elicited using the multiple bidding techniques was discussed. The predictions and predictors of WTP at the different bid levels were demonstrated. Concerns about the choice of estimates for criterion validity comparisons were raised. As discussed in earlier chapters, the analysis of hypothetical WTP data directly influences conclusions on criterion validity. This chapter explores the factors that influence WTP for one TMN from open ended and interval data. Following the two-stage bidding process discussed in chapter 7 and section 8.2.3, all the respondents who participated in the valuation process were invited to indicate the maximum WTP value that they were willing to pay for one TMN. The dependent variables considered are therefore the maximum WTP for one TMN and mean WTP from the different bid levels. The aim of this chapter is to demonstrate the alternate estimations of mean WTP from open ended and interval data, and the effect of these on criterion validity assessments and conclusions. The predictors of WTP from the open ended data will be determined, and the predictions of mean WTP from all the bid levels. The rest of the chapter is structured as follows: in section 9.2 the methods used in this analysis are presented beginning with a brief discussion on the open ended elicitation technique. The results of the different analysis are presented in section 9.3. In section 9.4 the results are discussed with the chapter concluded in section 9.5.

9.2 Methods

9.2.1 The Open ended method

In the reviews presented in chapters 4 and 5, authors primarily used mean estimates for criterion validity comparisons. The difficulties with open-ended CV responses for estimating mean WTP have long been recognised in the literature (Chern & Kaneko 2007; Johannesson et al. 1999; He et al. 2002). Critics of the open ended method argue that the question leads to highly skewed responses, which are often insensitive to changes in the quantity or quality of the valuation good (Damschroder et al. 2007; Reaves et al. 1999; Johannesson et al. 1999). Despite this, mean estimates remain the primary summary statistic used in criterion validity comparisons in the majority of the papers identified in the systematic review discussed in chapter 5. However, such comparisons could potentially lead to varying, sometimes incorrect conclusions on criterion validity.

Where open ended questions are used to elicit maximum WTP following a bidding process, the effect of the starting point bids cannot be ignored. Further, as has been demonstrated in the previous chapter, ignoring protest and selection biases can lead to incorrect estimates of WTP values. Even when selection bias is not observed, careful consideration is needed in the choice of the models for open ended WTP data, to allow for the possible non-normality of data.

9.2.2 Estimating Mean WTP from dichotomous choice WTP data

Mean and median summaries can also be estimated for WTP data obtained using dichotomous choice questions. In the malaria WTP study, a two-stage bidding technique was used to elicit WTP. As was discussed in previous chapters (3 and 7), the method has several positive attributes, rendering it ideal for use in CV surveys. While we cannot observe WTP directly using this method we infer that the respondent’s WTP is greater than a given bid (if the respondent says yes to it) or less than the bid (if the respondent says no to it) (Haab & McConnell 2002). These responses form the higher and lower boundaries of the respondent’s WTP and represent intervals within which the respondents’ true WTP lies.

The concerns with the analysis of hypothetical open ended and interval WTP data for criterion validity comparisons arise because:

1. Summary statistics (such as the mean) may be different across the elicitation formats and bid levels (even for the same sample).
2. The mean WTP estimates may be different using the OE question and the interval data (with the differences statistically significant);
3. The predicted mean WTP values from the different bid levels and the OE data might be different (with the differences statistically significant);
4. These differences in mean WTP estimates affect conclusions on criterion validity;

Further, where multiple elicitation mechanisms such as the bidding process are utilised, a range of summary statistics can be obtained. However, as discussed in chapter 6, even with the different possible summary statistics, only one estimate is presented for criterion validity comparisons. Often the justification for the choice of estimate used in comparisons with actual values is not provided. This limits further interrogation of conclusions on hypothetical bias. Using the empirical dataset presented in chapter 7, I estimate mean values from the open ended and interval data. I will also determine the predictors of WTP from the open ended question. Finally, predicted WTP values will be determined from the interval data. The different mean estimates will be compared with discussions highlighting the impact of the hypothetical estimate on criterion validity assessments and conclusions. The variables and specific analysis employed in this chapter are presented in the next section.

9.2.3 Independent variables

The data used in this analysis was discussed in chapter (7). All the independent variables used in this analysis were justified in the same chapter too. These were classified into the below clusters:

1. Respondent and household characteristics (gender, religion, caste, type of house, household size and number of children in the household, main earner in the household);
2. Socio-economic characteristics (education, employment and income);
3. Malaria variables (knowledge, exposure and experience with the disease, knowledge of and use of prevention methods);
4. Treated malaria net variables (current ownership and source of current net).

9.2.4 Dependent variables

The dependent variables for the analyses presented in this chapter are based on the discussions presented in the introductory sections. These are:

1. Maximum willingness to pay for one treated mosquito nets in cash or instalments (Max WTP). Following the valuation questions discussed in chapter 7, all the respondents were asked to state a maximum WTP value for one TMN. Figure 8.2 illustrates the bid path and maximum WTP elicitations.
2. Mean WTP from interval data. This was estimated from the bid functions.

Specific details of the analyses for each of the dependent variables are presented in the next section.

9.2.5 Data analysis

The analyses presented in this chapter were conducted in three stages:

(1) Descriptive analysis to summarise the maximum open-ended data. These involved determining mean WTP values for the overall sample as well as at the different bid paths. Comparisons of mean WTP estimates were done across the intervention groups and chi-square tests used to test for differences in key attributes.

(2) Bivariate analysis to determine the unadjusted relationship between each of the independent variables on the mean WTP.

(3) Multivariate analysis to investigate the adjusted relationships between the independent variables on the mean WTP;

(4) Predictions of mean WTP values from the interval data.

The analysis is conducted in two stages: (1) For the combined bids and (2) For the different bid paths. Bivariate and multivariate analyses were not conducted for the interval data as this was covered in the chapter 8.

9.2.5.1 Regression models

Modelling the maximum mean WTP from open ended WTP data

In the malaria WTP study discussed in chapter 7, maximum WTP was elicited using an open-ended question following the bidding process. This value was only elicited from the respondents who indicated that they were willing to buy nets in the filter question discussed in chapter 8 (first dependent variable). From the reduced sample[38] of 1,189, more than one-fifth (257) of the study respondents indicated that they were not willing to buy nets. These were not included in the valuation process and therefore the analyses presented in this chapter include only 932 respondents.

[38] From the total sample of 1,200, four respondents who indicated DK or N/A for all the WTP questions were dropped from the analysis. In addition, a further 7 respondents provided their WTP values but these were inconsistent (e.g. expressing a lower amount for the total WTP for all the household nets than their WTP for one TMN, suggesting inconsistencies in their preferences probably occasioned by a lack of understanding of the valuation process). These too were dropped from the analysis, reducing the study sample to 1,189 respondents.

As this analysis uses all the respondents who were willing to buy TMNs (n=932), there was no risk of selection bias and therefore the models did not account for this. However, as has been discussed earlier, open ended WTP data are prone to outliers which significantly skew the distribution. In particular, zero or near-zero values on one extreme and significantly large values on the upper end, are common. Such outliers affect the normality of the mean WTP data and specification of the models therefore calls for a consideration of this.

The generalized linear model (GLM) suggested by McCaulagh (1989) is used for this analysis. Such analysis permits the model to be constructed for response variables that are not normally distributed (Gelman & Hill 2006). An additional strength of the GLM estimators is that nonlinear least squares are generalized. Generalizing optimizes them for a non-linear regression model which is believed to contain homoskedastic additive errors (Dobson & Barnett 2008). The GLM model is also lauded because of the ease of interpretation of the results, especially when compared to log transformed values (Song et al. 2013). When the OE elicitation question is asked following one or more bids, a potentially significant bias that may affect the estimates is starting point bids. To test for the effect of the starting point bids, these are included in the estimated models.

Estimating Mean WTP from dichotomous choice WTP data

The point estimates (mean and median) from such data are derived by fitting special models as proposed by (Mitchell & Carson 1989). In estimating mean WTP from these data, I make assumptions about the bid amounts at which the probability of saying yes will be zero (upper limit of the integral) and the probability of saying no will be one. In the case of double bounded dichotomous data, the respondent’s WTP is bound by the first and second bids and this is estimated. When the single bounded question is asked, the respondent’s WTP lies between zero and the bid amount. When a follow up open ended question is asked following the single or double bounded dichotomous choice question, the stated maximum WTP offers a boundary. In the analysis presented in this chapter, a likelihood function based on interval data as suggested by Hanemann (1991) is considered most appropriate. I further assume that the WTP follows a normal distribution and is therefore only defined for non-negative values. The predicted mean WTP estimates are derived directly from the model estimates.

9.2.6 Model diagnostics

The linktest (Cameron & Trivedi 2009) was used to examine specification errors in the models. This test works by creating a variable of prediction and a second one, of the squared prediction and fitting the specified model with the two variables. When a regression model is well specified, the statistical significance of the variable of the squared significance should not be statistically significant[39]. Further, the Hosmer Lemeshow test was used to check for the goodness of fit of the models (Archer & Lemeshow 2006; Hosmer & Lemeshow 2013)[40].

9.3 Results

The results of the analysis are presented in this section. In the first part, the descriptive statistics are presented. These are followed by the model estimation results and the predicted mean WTP estimates from the interval data.

9.3.1 Descriptive statistics

The study sample was described elaborately in section 8.3.1 (chapter 8). The descriptive results in this chapter will focus on the responses to the open-ended question and the mean estimates from the different bid paths. Comparisons across the intervention groups for the specific variables will also be highlighted. Additional analyses by intervention group are presented in appendix 23.

[39] The linktest is based on the idea that if a model is properly specified no additional independent variables should be significant above chance. The link test looks for a specific type of specification error called a link error wherein a dependent variable needs to be transformed (linked) to accurately relate to independent variable. The link test adds the squared independent variable to the model and tests for significance versus the non-squared model. A model without a link error will have a nonsignificant t-test versus the unsquared version.

[40] The Hosmer Lemeshow test measures how well a model is specified by grouping cases together according to their predicted values from the logistic regression model. These predicted values are arranged from the lowest to the highest and separated into several groups of approximately equal size. For each group, the observed number of events and non-events is calculated, and the expected number of events (the sum of predicted probabilities for all the individuals in the group) and non-events (the group size less the expected number of events) too. The observed counts and expected counts are then compared using Pearson's chi-square. Low p-values (significance level, usually set at 0.05) suggest a poor fit of the model and thus it should be rejected while a high p-value suggests a good fit of the model.

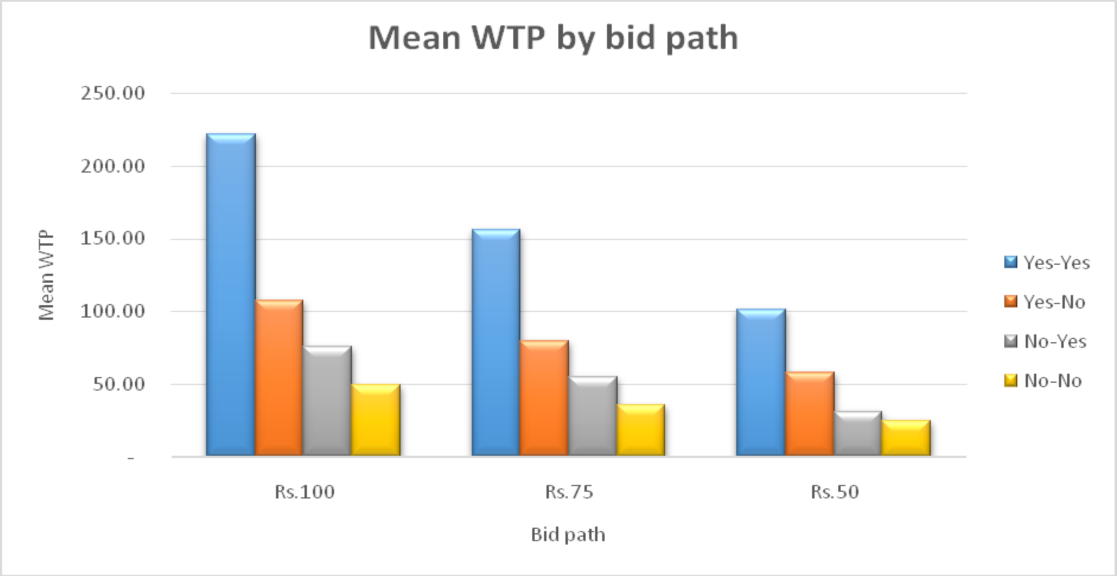
All the respondents who participated in the two-bid process were asked to state the maximum amount of money that they were willing to pay for one TMN. For each starting bid, there were four possible end points and mean summaries as detailed in table 9.1. For the 932 respondents, the mean WTP was Rs. 72.93 (SD 27. 54) with a range of Rs.15 – Rs. 300. Mean amounts were also determined for the twelve-different bid path end points. These results are further illustrated presented in table 9.1.

Table 9-1: Mean WTP estimates by bid path

Starting bid	**First bid response**	**Second bid response**	**Mean amount (Rs.)**
Rs. 100	Yes	Yes	221.8
Rs. 100	Yes	No	107.25
Rs. 100	No	Yes	75.9
Rs. 100	No	No	49.7
Rs. 75	Yes	Yes	156.2
Rs. 75	Yes	No	79.9
Rs. 75	No	Yes	55
Rs. 75	No	No	35.9
Rs. 50	Yes	Yes	101.7
Rs. 50	Yes	No	57.8
Rs. 50	No	Yes	31.4
Rs. 50	No	No	25

From the summary, it is evident that the higher starting point bids led to higher stated mean WTP amounts and vice versa. This is further illustrated in figure 9.1.

Figure 9-1: Mean WTP by starting bid path



At the stated maximum WTP, respondents were willing to purchase between one to eight nets with an average of two. More than two-thirds of the respondents would make a one-time payment for all the required nets with the remaining opting for instalment payment. More than half the respondents (54.73%) would purchase the TMNs in the months of May to October.

Less than half of the interviewed respondents (43.06%) would take the decision to buy the required nets for the household. Further, cash income would be used to purchase the TMNs for less than half (45.16%) of the study households.

In comparisons across the intervention groups, the mean WTP for TMNs is higher and similar for respondents in the ACT, IRS and OTA groups (Rs. 60+) but nearly half this amount (Rs.38) in the TMN group.

9.3.2 Model estimation results

9.3.2.1 Maximum WTP for one treated mosquito net

Both the base and reduced models were determined for the estimation of factors influencing the maximum WTP for one treated net. The results of the reduced models (table 9.2), which is more precise, are discussed in the next section. The base model output is presented in appendix 30. The results from the combined bids analysis are presented first and this is followed by the analysis of the separate bid paths.

The maximum WTP for one net increased by a factor of 2.6 for every net purchased by the household from the market. However, maximum WTP values were reduced for every unit increase in the household size, where respondents worked in animal husbandry or business, for households that practiced the Parsee faith, and those from the scheduled tribe. Maximum WTP values were also reduced for respondents who did not consider mosquitoes to be a nuisance and with every increase in the number of family members who suffered from malaria in the last month.

Results from the analysis of the separate bid paths highlight the stark differences in the predictors of maximum WTP (Tables 9.3 – 9.5). In particular, there were no similarities in the predictors of maximum WTP across the different bid paths. While the household caste (scheduled tribe) is a significant predictor of maximum WTP for both the Rs.50 and Rs.75 bid path with approximately similar magnitudes, the direction of effect is different.

Finally, in previous discussions, relationships between the starting point bids and the maximum WTP values were established. Respondents who accepted the first bids (Rs.50, Rs.75, and Rs.100) stated a maximum WTP amount that was 11.9% higher than those who rejected this bid. For those who accepted the second higher bid, the maximum WTP was 43.4% higher than those who rejected this bid. This is the greatest magnitude of effect among the three bids. Lastly, the maximum WTP value was 15.6% higher those who accepted the second lower bid than among those who did not. Similar results were obtained in the analysis of the separate bids. In these, the maximum WTP was highest for respondents who accepted the second higher bid. The magnitude of effect was lowest among respondents who accepted the second lower bid. This relationship maintained across the Rs.50 (table 9.3), Rs.75 (table 9.4) and Rs.100 starting bids (table 9.5).

Table 9-2: Model Outputs: Maximum Willingness to pay for nets _ Combined bids

Variable (Omitted category)	**Categories**	**Coefficient# (Robust standard error)**
No. of people living in the house		-0.935***(0.344)
Respondent's Main Occupation (Agriculture)	Animal Husbandry	-4.505*(2.629)
	Business	-3.687**(1.462)
	Labour Work	
	Others	
	Service	
Religion of the household (Christian)	Hindu	
	Muslim	
	Parsee	-12.11***(3.484)
Caste household belongs to (Scheduled caste)	Other backward caste	
	Scheduled tribe	-5.523**(2.174)
Whether respondent considers mosquitoes to be a nuisance (Major nuisance)	Minor nuisance	
	No Nuisance	-6.961*(4.135)
No. of nets purchased from market		2.626**(1.202)
No. of family members suffering from malaria last month		-2.732**(1.189)
Willingness to buy at first bid (Yes)		11.90***(1.745)
Willingness to buy at Second higher bid (Yes)		43.46***(6.144)
Willingness to buy at Second lower bid (Yes)		15.60***(1.905)
Constant		59.14***(3.090)
Pseudo R^2		0.1324
Linktest		0.00241[a]
Observations		932

#The estimated parameters and asterisks show significance level of 1% (***), 5% (**) and 10% (*)

[a] p=0.068

Table 9-3: Model Outputs: Maximum Willingness to pay for nets _ Rs.50 bid path

Dependent Variable 5: Maximum Willingness to pay for one net	**Categories**	**Coefficient[#] (Robust standard error)**
No. of people living in the house		-0.605**(0.248)
Main earner's education (Graduation)	Illiterate	-13.24**(5.759)
	Primary	-15.28**(5.953)
	Secondary	-13.18**(6.108)
Religion of the household (Christian)	Hindu	
	Muslim	
	Parsee	14.94***(2.899)
Caste household belongs to (Scheduled caste)	Other backward caste	6.608**(3.341)
	Scheduled tribe	5.380**(2.211)
No. of nets purchased from market		2.579***(0.827)
Willingness to buy at first bid (Yes)		26.00***(1.811)
Willingness to buy at Second higher bid (Yes)		42.34***(2.292)
Willingness to buy at Second lower bid (Yes)		5.931**(2.673)
Constant		36.60***(5.775)
Pseudo R^2		0.1793
Linktest		0.016[a]
Observations		321

[#]The estimated parameters and asterisks show significance level of 1% (***), 5% (**) and 10% (*)

[a] p=0.166

Table 9-4: Model Outputs: Maximum Willingness to pay for nets _ Rs.75 bid path

Variable (Base)	**Categories**	**Coefficient[#] (Robust standard error)**
Interview village (Treated mosquito nets)	Active Case Detection	4.404**(1.982)
	In-house spray village	5.165**(2.063)
	Outside trial area	5.278***(1.865)
Respondent’s Main Occupation (Agriculture)	Animal Husbandry	-2.283*(1.239)
	Business	-4.052***(1.509)
	Labour Work	
	Others	
	Service	
Caste household belongs to (Scheduled caste)	Other backward caste	
	Scheduled tribe	-5.526***(1.731)
Type of house (Kaccha)	Pucca	
	Semi pucca	
Total household disposable income (log)		-0.246**(0.0987)
Mosquito net among preferred methods (Yes))		2.397**(1.050)
No. of nets owned		0.865*(0.525)
If any disease is caused by mosquito bites (No)		3.385**(1.581)
Willingness to buy at first bid (Yes)		25.16***(1.056)
Willingness to buy at Second higher bid (Yes)		74.16***(5.367)
Willingness to buy at Second lower bid (Yes)		19.37***(3.352)
Constant		31.42***(3.761)
Pseudo R^2		0.2556
Linktest		0.00021[a]
Observations		308

[#]The estimated parameters and asterisks show significance level of 1% (***), 5% (**) and 10% (*)

[a] p=0.673

Table 9-5: Model Outputs: Maximum Willingness to pay for nets _ Rs.100 bid path

Variable (Base)	Categories	Coefficient[#] (Robust standard error)
Religion of the household (Christian)	Hindu	
	Muslim	-14.85***(2.555
	Parsee	
Whether respondent considers mosquitoes to be a nuisance (Major nuisance)	Minor nuisance	-1.797*(0.992)
	No Nuisance	2.084*(1.214)
Total no. of prevention methods used		2.424**(1.152)
No. of family members suffering from malaria last month		-4.222***(1.604)
Expenditure incurred on malaria treatment (log)		0.686*(0.390)
Willingness to buy at first bid (Yes)		31.19***(1.310)
Willingness to buy at Second higher bid (Yes)		114.0***(13.19)
Willingness to buy at Second lower bid (Yes)		25.72***(1.182)
Constant		-14.85***(2.555
Pseudo R^2		0.1821
Linktest		0.00007[a]
Observations		303

[#]The estimated parameters and asterisks show significance level of 1% (***), 5% (**) and 10% (*)

[a] p=0.774

9.3.2.2 Estimation of mean WTP from dichotomous choice data

The variables which are expected to influence WTP at the different bid points were discussed in chapter 8. The predicted mean WTP values for both the combined and separate bid paths are presented in table 9.3 with a discussion in section 9.4. Across the different models, the mean stated maximum WTP values are highest at the second higher bid and lowest at the second lower bid. This accords with apriori expectations. Further, as previously pointed out, a relationship between the starting point bid and the stated maximum WTP values is identified. Specifically, respondents who were offered the highest starting point bid stated the highest maximum WTP values and vice versa. This further confirms the multivariate regression results that starting point bids in a bidding elicitation format as is the case with the current study positively predicts maximum WTP values.

Table 9-6: Predicted mean values from the open ended and interval data

Estimate level	**Combined sample**	**Rs.50 Starting bid**	**Rs.75 Starting bid**	**Rs.100 Starting bid**
Maximum WTP (all sample)	72.93 *(0.56*)* *(71.89 – 74.05#)*	63.03 *(1.08*)* *(60.85– 65.21#)*	75.64 *(1.11*)* *(70.45 – 74.83#)*	83.72 *(1.82*)* *(80.14 – 87.31#)*
First bid	80.79 *(0.63*)* *(79.54 – 82.04#)*	66.34 *(1.04*)* *(64.28 – 68.39#)*	82.90 *(1.06*)* *(80.81 – 84.99#)*	115.74 *(2.91*)* *(109.96 –121.51#)*
Second higher bid	121.18 *(0.69*)* *(119.79 – 122.56#)*	101.78 (*(0.67*)* *(100.43 – 103.13#)*	156.25 *(1.32*)* *153.10 – 159.39#)*	221.875 *(1.195*)* *219.04 – 224.70#)*
Second lower bid	62.93 *(0.28*)* *(62.36 – 63.49#)*	31.42 *(0.52*)* *(30.35 – 32.50#)*	55.05 *(0.30*)* *(54.44 – 55.66#)*	75.95 *(0.19*)* *(75.55 – 76.34#)*

*Standard error; #95% CI

9.4 Discussion of results

The factors influencing the stated WTP value when dichotomous choice methods are used were discussed in the previous chapter (8) and will therefore not be repeated. From the analysis of factors influencing the maximum WTP for one TMN, most of the independent variables predicted in accordance to the empirical evidence and theoretical basis outlined in chapter 7. This further confirms the evidence on the predictors of WTP for a TMN. The GLM model permitted the use of the entire WTP data with a good fit for the specification.

Contrary to the apriori expectation, the interview village was not a significant determinant of maximum WTP, and this was the case even when the Rs.50 and Rs.100 bid paths were analysed separately. Notably also, unlike in previous models, whether the respondent was a main earner or not (which would determine whether they were decision makers or budget holders), was not a significant determinant of WTP values. However, it is clear that respondents were thinking about their budgets and the effect of malaria episodes in responding to the valuation question, as shown in the combined bids analysis. This is evidenced by the magnitude of influence in these variables. Of significant note also is the fact that household income does not seem to be a predictor for maximum WTP. This is contrary to apriori expectations.

As indicated earlier, the finding on starting point bias confirms existing evidence (Vossler et al., 2003), further indicating a good specification of the model. This further advances the argument for careful design in the bidding process. In most of

CV-WTP studies using the bidding method, there is limited discussion on the choice of starting point bids. In the malaria WTP study, qualitative interviews were conducted to inform the construction of the hypothetical market. However, the bids were randomly picked (Bhatia & Fox-Rushby 2002). The nature of the goods for which a CV study is used is such that there are no price signals, making it difficult to determine an appropriate bid price. However, the inclusion of the bids as part of the tool pre-test process might help construct appropriate starting point bids. Even then, the use of multiple bids as is the case in the current study allows for variation as has been suggested in the literature (Bateman et al., 2002). Further, combining the bidding process with a final open-ended question allows for the measurement of the full consumer surplus (Mitchell & Carson 1989).

The magnitude of the influence of the starting point bid on stated WTP is also worth further consideration. As was noted earlier, the greatest influence is on the second higher bid, and this is followed by the first bid. Both of these then provide some signals for where the bids might be adjusted to minimise their impact on stated WTP amounts. Overall, that the majority of the independent variables investigated accords with the economic literature and empirical evidence is a good sign of the careful design and execution of this study. The analysis of the data points to the varied possible hypothetical WTP estimates even for the same population, providing further justification for the call for a critical assessment of criterion validity assessments.
The predicted mean values are different across the bids and much lower than the mean WTP values derived by averaging the stated values. This further demonstrates the differences in the estimates.

As noted in the previous chapter, this analysis was limited to the available dataset, and the variables as specified in the dataset. A primary study might have permitted the exploration of the additional elicitation formats on the assessment and conclusions on criterion validity. However, the secondary dataset utilised was considered sufficient for the purposes of the current analyses. The justification and advantages of using a secondary dataset were presented in previous sections. Additional investigations of the effect of different elicitation formats on WTP estimates are therefore recommended for the future. This will generate further evidence, possibly leading to firm conclusions on the subject.

In addition, specified differently, the model estimation results might have been different. These too could be explored in future research. For instance, the analysis of open ended data presented here utilised the GLM estimator. The GLM permits the use of the full range of values, including zeros. This was considered appropriate in this analysis for that reason. Further, as the study was conducted in a low-income setting, it is likely that zero values truly reflected the respondents' willingness to pay for the TMN. The same is true for income levels. The GLM would take into account the entire distributions of income. The ability of the GLM to capture societal values in this way renders it most ideal for this specific analysis. Other models that may have been used include those that truncate values at a value such as zero. These were considered unsuitable for this analysis.

Further, as there is no consensus on the effects of different experimental protocol on WTP, independent variables were included in the models largely in an exploratory way. It is likely that a different set of independent variables would have elicited different results. However, as discussed earlier, the analyses presented here was limited to the variables available in the secondary dataset. Further the model diagnosis results confirmed that the model was well specified.

9.5 Implications for criterion validity

In the majority of the literature where open ended WTP elicitation methods are used, mean (and median) summaries are presented for criterion validity comparisons. This is the case too where multiple methods, as was the case with the empirical dataset utilised for this analysis are presented. Mean estimates obtained from the open-ended data or percentages of responses to different bids where discrete choice data is used are presented. Such summaries are then compared with actual values obtained using one of the elicitation techniques discussed. The method used in the actual survey could be one of the methods used in the hypothetical survey, or totally different. However, as has been demonstrated in the present and previous chapter, the factors which influence WTP for an ITN differ, depending on the elicitation technique utilised. Differences are also evident even with multiple elicitation points using the same technique (e.g. multiple bids). The current chapter focussed on modelling the open-ended data while estimates of mean and median WTP were derived for the dichotomous choice data. The analysis demonstrated that:

a. Average summaries presented from open ended data differ significantly from predicted mean values. While the average summaries are informative, they are unadjusted and therefore not the best reflections of real estimates.

b. The predicted mean values are significantly different for the different WTP elicitation bids. In addition, these differ across the elicitation methods.

c. The predictors of WTP for the open-ended data differ from those identified with the discrete choice data.

d. Both the predicted mean values and predictors of maximum WTP are different across different bid paths.

These differences affect criterion validity assessments and conclusions thereof. As argued in the previous chapter, criterion validity assessments presenting the full range of estimates are likely to be more accurate than those based on aggregated estimates. Further, presenting the full range of estimates allows for criterion validity assessments at multiple levels. However, with no guidelines on the conduct and reporting of criterion validity assessments, authors have often decided on the estimates to present and use for such assessments. The justification for the selected estimates is often not provided. This potentially introduces subjectivity and bias, questioning criterion validity thereof conclusions thereof. Researchers continue to explore the effect of different experimental protocol on hypothetical bias and this in varied ways. The limited number of criterion validity assessments is also an additional hurdle in efforts to improve the method.

9.6 Conclusion and chapter summary

As demonstrated in this chapter, the analysis of hypothetical WTP values significantly influences criterion validity assessments and the conclusions derived from these assessments. In the present chapter, the alternate estimates of mean WTP that can be obtained from the same sample have been illustrated. The effect of these on criterion validity assessments and conclusions has also been discussed. Evidently, the variety in the assessment methods, lack of consensus on the choice of estimate for criterion validity assessments, and inconsistencies in reporting of WTP data contribute to current conclusions on the criterion validity of CV-WTP. These in turn add to the ongoing criticism of the CV-WTP method. However, the method still offers great potential for benefit assessment. In the final chapter, the findings from the thesis are summarised and recommendations made for advancing the discussion on the criterion validity of CV-WTP methods.

Chapter 10 Conclusions

The uptake of contingent valuation studies in benefit valuation has been very slow, particularly in the health sector. This is despite the theoretical strengths of the method in the valuation of non-market goods. The hesitation with the use of the method partly arises from the perceived complexities in designing and executing CV-WTP studies. But, study design issues are overshadowed by the overwhelming critique of the method. As has been discussed in the thesis, concerns have been raised about the validity of the method, with critiques arguing that hypothetical WTP values are poor estimates of actual value. The premise therefore is that hypothetical WTP values are not credible signals of actual value and should not be used in decision making. However, in the analyses presented in this thesis, it was established that the evidence on criterion validity is more mixed than authors are reporting and similarly, the magnitude of hypothetical bias is not as significant as presented in the literature.

Previous reviews of empirical assessments of criterion validity highlight the effect of elicitation formats on hypothetical bias. The majority of criterion validity assessments where multiple elicitation techniques are used to obtain hypothetical WTP estimates present an aggregated estimate for comparisons with actual values. However, multiple elicitation techniques, and valuations at multiple points not only lead to different estimates, but the predictors of WTP are different. When these differences are significant, aggregating such values might lead to incorrect conclusions on the criterion validity of CV-WTP. The analyses presented in this thesis demonstrate this point.

The majority of criterion validity assessments are based on comparisons of summary hypothetical WTP and actual values. However, criterion validity can also be assessed through an investigation of the predictors, and predictions of hypothetical WTP. These are then compared with findings from surveys of actual values. In advancing the discussion on the methodological issues with the conduct of criterion validity assessments, this was explored in this thesis. In this concluding chapter, the sections are structured as follows: in 10.1 the contributions of the thesis to literature on the criterion validity of CV-WTP are summarised. In section 10.2, suggested guidelines on the reporting of CV-WTP criterion validity assessments are presented;

the limitations of the thesis are presented in section 10.3 and the implications for research are outlined in section 10.4. Some recommendations for future research are detailed in section 10.5 with some final comments concluding the chapter and thesis in section 10.6.

10.1 Contributions of the thesis

This thesis utilised a variety of methods to address the question of the criterion validity of CV-WTP methods. This was done through the conduct of four systematic reviews, and three distinct but related empirical analyses. I therefore contributed to the growing body of knowledge on the criterion validity of contingent valuation WTP methods in the ways detailed below.

In chapter 3, I synthesised the theoretical framework within which contingent valuation studies are conducted. In doing this, the strengths of the contingent valuation method in assessing non-market benefits were demonstrated. As discussed in the chapter, the importance of such estimates for cost benefit analysis is evident. Yet, the use of the contingent valuation method is limited, especially in the health sector. Concerns with the use of the CV method relate primarily to the validity of the hypothetical WTP estimates. With its origins in measurement theory, the measurement of validity is a cyclic process. As one researcher observes, as there are many hypotheses that can be tested, validation is a process with validity as the outcome (Streiner et al. 2008). This chapter contributed to the thesis by providing a theoretical framework within which subsequent discussions and analysis of CV-WTP are situated.

In chapter 4, I provided a critical appraisal of the evidence on the methods used to assess the different types of validity. This was done by systematically reviewing empirical studies assessing the different types of the validity of CV-WTP. This chapter demonstrated that there is a gap in criterion validity research. The chapter also established that there is variety and possibly confusion in the terms used to describe validity. The most common form of validity tested is the construct or theoretical validity. Criterion validity, also referred to as external validity in the literature, is primarily assessed by comparing hypothetical WTP estimates obtained through surveys and laboratory experiments with observations of real market transactions in simulated market experiments. Of primary concern to this thesis, the review highlights the relatively limited number of criterion validity assessments of CV-WTP overall, and particularly in health. This chapter therefore contributed to the thesis by demonstrating the gap in evidence on criterion validity.

Chapter 5 contributes to the thesis by critically analysing the evidence on criterion validity. This is done through the conduct of two systematic reviews. In the first, the methods used to investigate the criterion validity of CV-WTP and the conclusions thereof are evaluated. This is done through a review of reviews that have synthesised the evidence on criterion validity across the sectors. This chapter further demonstrated the variety in the methods that have been used to investigate criterion validity. It also shows that there is no consensus on the most appropriate method for use in assessing criterion validity. While the conclusions on the presence of hypothetical bias in CV-WTP studies are similar across the reviews, the evidence on the magnitude is different.

Guided by the methods and evidence from the review of reviews, a focussed review on empirical studies assessing criterion validity was conducted. The results of this review were presented in chapter 5 too. As with the summary of reviews, the variability in criterion validity assessment methods was demonstrated. The review also established the confusion in the terms used to denote criterion validity assessments. These potentially make the identification of empirical assessments difficult, and thereby limiting opportunities for the synthesis of the evidence. The reviews also highlight the dearth in criterion validity assessments, particularly in the health sector. The limited evidence base means that firm conclusions on the methods used to assess criterion validity, and the evidence thereof are not robust. This review presents a critical assessment of the current evidence on the criterion validity of CV-WTP methods.

Following the synthesis of the evidence on the criterion validity of CV-WTP methods in the previous chapter, in chapter 6, the magnitude of hypothetical bias is quantified. The last published evidence on the magnitude of hypothetical bias was conducted more than a decade ago (2005). To date, a total of three meta-analysis of criterion validity have been conducted. The meta-analysis presented in this chapter is the first to use the strict criteria of including only studies involving a financial transaction (market prices or related prices) in the actual survey. As discussed in chapter 3, market prices are considered a valid criterion for validity assessments. As such, the magnitude of hypothetical bias in CV-WTP studies reported in this chapter could be regarded as the most current evidence across the sectors.

The meta-analysis also highlights challenges with synthesising data from criterion validity assessments. The variety in the type and depth of reporting of criterion validity assessments is noted. As a result of this variety, studies reporting percentage and mean summaries were synthesised separately. The magnitude of hypothetical bias from the reported studies is 1.785 for studies reporting mean summaries and 2.327 for studies reporting percentage summaries. Previous meta-analyses reported mean values ranging between 0.80 and more than 3. Notably, previous meta-analysis do not differentiate between studies reporting mean and percentage summaries. The variety in the reporting of estimates further limits the use of all the available empirical assessments of criterion validity for such meta-analyses. The results of the meta-regression reported in this chapter identify a range of experimental protocol that influences hypothetical bias. Key among these is the effect of the WTP elicitation techniques. These may be explored further in empirical assessments of criterion validity. The analysis further suggests that hypothetical bias is significantly larger with open ended WTP elicitation methods and less with discrete choice methods such as the dichotomous choice methods. Relatively few studies have explored the effect of the analytical methods used with multiple hypothetical WTP elicitation techniques on criterion validity assessments and conclusions thereof.

The systematic review presented in chapter 5 identified very few empirical assessments of the criterion validity of CV-WTP in the health sector. In chapter 7, the process used to identify a suitable empirical dataset for use in further investigating the effect of different experimental protocol on the criterion validity of CV-WTP was discussed. A systematic search, following established criteria was conducted. For the investigation of the effect of the analytical methods on criterion validity, a dataset that employed multiple elicitation techniques to elicit hypothetical WTP values was needed. A suitable dataset needed to have been conducted in the health sector and reported the full range of estimates from the hypothetical survey. In addition, the dataset needed to have collected data on and reported various socio-economic characteristics. This follows recommendations by Arrow et al (1993) on a good CV-WTP study. By applying the criteria to the empirical studies identified during the systematic review of the criterion validity of CV-WTP, the Malaria WTP study was identified. The study was conducted as part of a large randomised clinical trial on Malaria control interventions in Surat, India. In the hypothetical survey, a multiple bidding process followed by an open-ended question was used to elicit WTP values. A range of respondent socio-economic characteristics were also obtained.

In the same chapter, a systematic review conducted to identify the factors which influence WTP for treated mosquito nets (TMN) is presented. The purpose of this review was to determine a range of independent variables for use in specifying regression models in subsequent analysis. The evidence identified on a range of variables was clustered into four broad categories: household background characteristics, malaria prevention measures, net ownership variables and variables related to knowledge of malaria (disease) and expenditure related to management of the disease.

Using the Malaria WTP dataset, chapter 8 presents an empirical analysis of WTP data elicited through discrete choice techniques. In the Malaria WTP study, hypothetical WTP was first obtained using a multiple stage bidding technique. Prior to the valuation question, a single bounded dichotomous choice question was asked to filter respondents who were in the market for mosquito nets from those who were not. Both the filter question and the multiple bidding technique presented points from which hypothetical WTP could be estimated. The analysis presented in this chapter demonstrates the different summary estimates that can be obtained from the descriptive analysis of discrete choice data. The chapter also illustrates the effect of estimate choice on criterion validity comparisons and conclusions. Criterion validity assessments primarily present a single estimate for comparisons with actual values. This is the case even where multiple elicitation formats or multiple bid levels are used. This trend was established in the systematic reviews presented in chapters 4 and 5.

However, the analysis in chapter 8 further demonstrates that the predictors of WTP at the different bid levels are in fact different. Further analysis establishes that the predicted WTP values will be different too. The chapter therefore questions the use of aggregated summaries from such multiple bid levels as comparators for criterion validity assessments. The analyses also demonstrate that multiple WTP predictions can be obtained from such bidding techniques. With authors presenting only one of these estimates for criterion validity comparisons, the presented conclusions on criterion validity are not entirely accurate. The results obtained from the analysis presented in this chapter advance the argument for conducting and reporting criterion validity comparisons from the multiple estimations.

Chapter 9 presents an analysis of open ended and the interval data elicited following multiple stage bidding formats. The open-ended question was asked following the bidding process discussed in chapter 8. Given the descriptive results and predictions of mean WTP based on the adjusted predictors, the chapter illustrates the multiple estimates which can be obtained from such open ended and interval data. Notably, the chapter demonstrates that the predictions of mean WTP differ across bidding levels and elicitation methods. Chapter 9 further questions the aggregation of estimates from multiple elicitation formats, and bid levels, for comparisons with actual values. Criterion validity assessments based on such aggregated values would lead to incorrect conclusions on criterion validity. The analysis presented in this chapter further demonstrate the need to investigate the methods used to analyse WTP data, and the estimates presented for criterion validity assessments.

Based on all the analyses and discussions, a primary output of thesis is initial guidelines for the reporting of criterion validity assessments in CV-WTP studies. These are summarised into a checklist provided in table 10.1. The suggestions build on earlier guidelines for the conduct of economic evaluations as discussed in the following section.

10.2 Guidelines for the reporting of criterion validity assessments

The challenges associated with the reporting of economic evaluations of health interventions are widely acknowledged (BMJ 2013). This has led to the development of guidelines for the assessment and reporting of such evaluations. The CHEERS statement is one of these and offers consolidated guidelines for reporting of economic evaluations of health interventions (*ibid*). In his classical text, Drummond classified the guidelines for reporting of economic evaluations according to three purposes as: (1) those linked to a formal requirement for reimbursement purposes; (2) guidelines related to ethical standards and; (3) guidelines which are related to the maintenance of and advancement of methodological standards. The last set of guidelines are supposed to aid in the interpretation of economic evaluation results and thus aid in decision making. Drummond also outlined a checklist for the critical appraisal of published economic evaluations of health care programmes which he suggests could be used to enhance the structure and quality of economic evaluation study reports (Drummond et al., 1997).

Using Drummond's classification, Hjelmgren categorised guidelines identified from a systematic review on the reporting of economic evaluations (Hjelmgren et al. 2001). The review identified eleven guidelines issued in Europe, North America, and Australia which are aimed at improving economic evaluations methods in health. Six of the guidelines are primarily related to the pharmaceutical industry with two each covering medicine and general economic evaluations while one covers health technology assessments. Seven of these guidelines include CBA as the type of analysis considered and it is the most preferred method in only one of these[41]. The use of contingent valuation methods is preferred for the assignment of values to outcomes in the CBA in these guidelines. The authors provide a case for the use of WTP studies, including suggestions for the use of dichotomous choice questions and discussions on who should be asked WTP questions. While the authors highlight the issues with the reliability and validity of WTP methods, the only suggestion offered for the assessment of validity is the incorporation of scope tests to assess the congruence of the direction of WTP responses with the benefit assessed in the study.

Evidently, significant work has been conducted on guidelines for the conduct of economic evaluations, including contingent valuation studies (Mitchell & Carson, 1989; Bateman et al, 2002). However, missing in the literature are guidelines for the assessment and reporting of criterion validity in CV-WTP studies. This poses significant challenges for researchers attempting to systematically synthesise the available evidence for decision making. The lack of specific guidelines for the conduct and reporting of criterion validity assessments also presents challenges for journal editors and reviewers, as has been highlighted for other economic evaluations (Rennie & Luft 2000). The lack of guidelines could potentially lead to

[41] Canadian Coordinating Office for Health Technology Assessment. Guidelines for Economic Evaluation of Pharmaceuticals (2nd edn.). Ottawa: Canadian Coordinating Office for Health Technology Assessment (CCOHTA), 1997.

biases in the publication of criterion validity studies, further hindering the methodological research. This therefore justifies the call for some quality assurance mechanisms, as guidelines are expected to provide (Neumann et al. 2000; Sanders et al. 2016; Rosen et al. 2005).

Previous systematic reviews on the criterion validity of CV-WTP studies have highlighted the difficulties in pooling estimates from criterion validity assessments. Among the highlighted challenges include incomplete or missing estimates, a general lack of justification for the choice of analytical techniques, lack of clarity on the basis for decisions on confirmation (or not) of criterion validity. The eventual effect of having such missing or incomplete data hinders the pooling of estimates from criterion validity assessments, which would permit the determination of an overall magnitude of the effect of hypothetical bias on stated WTP values.

In addition, the failure to justify the processes limits the understanding and inferences one can make from such datasets. However, the challenges with the design and execution of criterion validity assessments are acknowledged. The suggestion to develop some guidelines for the assessment and reporting of criterion validity assessments are early attempts to harmonise studies. Hopefully, this can lead to a sufficient scope of literature that would permit further investigations into the effect of different design and analytical attributes on criterion validity assessments and the conclusions thereof.

Previous systematic reviews on the criterion validity of CV-WTP studies have highlighted the difficulties in pooling estimates from criterion validity assessments. Among the highlighted challenges include incomplete or missing estimates, a general lack of justification for the choice of analytical techniques, lack of clarity on the basis for decisions on confirmation (or not) of criterion validity. Such missing or incomplete data hinders the pooling of estimates from criterion validity assessments, which would permit for the determination of an overall magnitude of the effect of hypothetical bias on stated WTP values.

In addition, the failure to justify the processes limits the understanding and inferences one can make from such datasets. However, the challenges with the design and execution of criterion validity assessments are acknowledged. The suggestion to develop some guidelines for the assessment and reporting of criterion validity assessments are early attempts to harmonise studies. Hopefully, this can lead to a sufficient scope of literature that would permit further investigations into the effect of different design and analytical attributes on criterion validity assessments and the conclusions thereof. As more studies are conducted, and publications made available, the importance of clarity in reporting should be an essential aspect of the research and evidence generation process. The opportunity cost from decisions based on incorrectly reported findings or lack of transparency is significant. In particular, current debates on the criterion validity of CV-WTP may not be based on a correct understanding of the magnitude of hypothetical bias (Sanders et al. 2016). As has been demonstrated in the analysis presented in this thesis, it is not the method, rather, the methodological issues that warrant further investigation. In remedying the methodological issues related to the conduct and reporting of criterion validity assessments, the method will be improved.

In order to develop guidelines for the assessment of the criterion validity of CV-WTP, a sufficient body of evidence on the different experimental protocol must be established through empirical research. Following the empirical work, it must be possible to pool these results in an informative way. Hence, there is need for guidelines to standardize the reporting of criterion validity assessments. In early attempts to contribute to this discussion, I provide some suggestions for the reporting of criterion validity assessments. These are specific for criterion validity assessments in which hypothetical WTP values are compared with actual values obtained through a simulated market experiment (SME). This is because as discussed in chapter 4, market prices are regarded as the best criterion for validity assessment. In CV-WTP criterion validity assessments, market prices are obtained through SMEs. The suggested guidelines are based on the discussions and analysis presented in this thesis. Where possible, each suggestion will be clearly linked to the relevant chapter in the thesis. Some of the suggestions are included to aid in generating further evidence on the effect of varied experimental protocol on the criterion validity of CV-WTP. The recommendations summarised by Drummond (1997) as guidelines common to many of the existing economic evaluation formats still apply. The suggestions for the reporting of criterion validity assessments are classified into four broad categories:

1) Overview of the CV study largely borrowed from earlier guidelines on the conduct of CV-WTP studies (Mitchell & Carson, 1989; Bateman et al, 2002).

2) Criterion validity assessment attributes

3) Study and results

4) Conclusions on criterion validity and declarations

Specific elements of each of these categories are provided below and the summary checklist outlined in table 10.1.

A) Overview of the study

1. The study or report should be clearly labelled as an assessment of the criterion (or external)[42] validity of CV-WTP methods. The systematic reviews presented in chapters 4 and 5 highlighted the confusion in the terms used to define validity. This likely contributed to the challenges in identifying such studies, as discussed in the relevant chapters and a later section. Harmonization of the definitions will make identification of studies easier.

2. The purpose of the study is clearly indicated. For example, is the assessment conducted to test experimental protocol?

3. The hypothetical and actual (SME) study should be clearly described.

4. The valuation good and the sector should be defined. The purpose of this is to aid in comparisons and further test experimental protocol.

5. The elicitation formats used in both the hypothetical CV study and the actual (SME) survey should be clearly indicated. Testing of these will build the evidence base on the effect of elicitation methods on the criterion validity of CV-WTP methods.

6. The sample sizes, types and selection of samples for both the hypothetical WTP and the actual (SME) surveys should be reported. This is to help in generating evidence on the effect of varied experimental protocol.

[42] While I recognise the current variety in the terms used to define validity, I use the definitions detailed in appendix 3. These are drawn from measurement theory and psychology (Carmines & Zeller 1979).

7. The administration modes for both the hypothetical and actual (SME) surveys should be reported. This is to help in generating evidence on the effect of varied experimental protocol.

8. Studies should report both the overall study response rate and the response rate related to the valuation questions. This aids in the understanding and interpretation of the obtained estimates. For instance, lower response rates to the valuation question, compared to the overall study response rate might indicate protest responses which need to be acknowledged or investigated in the analysis of WTP estimates.

B) Criterion validity assessment elements

1. The methods used to assess criterion validity should be clearly outlined. Evidence from the reviews presented in chapters 4 and 5 further demonstrates the variety in the methods used to assess criterion validity. As discussed earlier, there is no consensus on the most ideal assessment method.

2. Whether a between-sample or within-sample analysis is used should clear. This is to help in generating evidence on the effect of varied experimental protocol.

3. The duration between the hypothetical CV survey and the SME or actual survey should be indicated. This is to help in generating evidence on the effect of varied experimental protocol.

4. Studies should provide information about the values presented in both the hypothetical and actual or SME survey (Value cues), where a relevant elicitation format is used. For instance, whether the value presented in the SME relates to estimates obtained from the hypothetical WTP survey.

5. The analysis methods, the basis of which comparators for criterion validity assessments are derived should be justified. The analysis presented in chapters 8 and 9 demonstrates that the multiple adjusted and unadjusted estimates can be obtained from the analysis of WTP data.

6. The choice of estimates (comparators) for use in comparisons with actual values for criterion validity assessments should be provided and justified.

C) Study results

1. Studies should provide details of the background characteristics of the respondents. This will help in further testing the drivers of hypothetical bias.
2. Studies should report the full range of estimates obtained (e.g. summary estimate and related measures of variability in the data). This aids in the understanding and interpretation of the entire dataset. Reporting all relevant estimates also aids in pooling of studies for a combined measure of effect. Attempts to quantify the magnitude of hypothetical bias for all the studies identified in the systematic review reported in chapter 5 were hampered by missing estimates from some of the studies
3. Estimates should be reported for all endpoints where multiple study attributes are evaluated (e.g. multiple study administration techniques, elicitation and analysis methods)
4. Where multiple methods (such as a dichotomous choice with follow up open ended) or multiple estimation techniques (such as a multiple level bidding method) are used to obtain hypothetical WTP values, disaggregated results should be presented. A key premise in the discussions in this thesis is that aggregated data limits the interpretation of study results and can be misleading in criterion validity assessments. This was evidenced in chapters 8 and 9.
5. Where multiple methods (such as a dichotomous choice with follow up open ended) or multiple estimation techniques (such as a multiple level bidding method) are used to obtain hypothetical WTP values, criterion validity assessments should be conducted for each hypothetical WTP summary. In addition to providing a robust assessment of criterion validity, this will help in identifying opportunities for further methodological research on criterion validity assessments.
6. Where possible, both the predictors and predictions of WTP should be assessed in hypothetical WTP surveys. As these estimates are adjusted for a

range of confounders, they are deemed more accurate than unadjusted estimates and might present better comparators for criterion validity assessments.

D) Conclusions on criterion validity and declarations

1. The comparisons done for the assessment of criterion validity should be clearly indicated. For example, whether the mean hypothetical WTP is compared with the mean value obtained from the actual survey.
2. Decisions on the validity (or lack of it) should be justified. While there is no agreement in the literature on acceptability margins for the confirmation of criterion validity, providing a justification for the decisions on criterion validity enhances the transparency of the process. This also provides an opportunity for further interrogation of the criterion validity assessment methods.
3. Authors should declare any conflicts of interests relevant for the study e.g. affiliations and sources of funding for the study.

Table 10-1: Checklist for the reporting of criterion validity assessments of CV-WTP studies

#	Checklist item	Yes	No	Partial	NA
	A. Overview of the Criterion Validity Study				
1	Identifies the study or report clearly as an assessment of criterion or external validity of WTP methods				
2	States the purpose of the study (e.g. testing experimental protocol, eliciting values for CBA)				
3	Describes the hypothetical survey setting.				
4	Describes the setting within which actual prices are obtained (e.g. SME).				
5	Describes the valuation good(s): Provision (whether public or private) Purpose of the good				
6	Elicitation format (s) indicated: Hypothetical survey Actual survey (SME)				
7	Hypothetical Survey Sample characteristics: Sample Size Sample Selection (e.g. Random, Purposive, etc.) Sample Type (e.g. Users / Non-Users, Students / Non-Students etc.)				
8	Actual Survey (SME) Sample characteristics: Sample Size Sample Selection (e.g. Random, Purposive, etc.) Sample Type (e.g. Users / Non-Users, Students / Non-Students etc.)				
9	Survey Administration Mode(s) indicated: Hypothetical survey Actual survey (SME)				
10	Analytical method (s) indicated: Hypothetical survey WTP values				
	Actual survey (SME) values				

#	Checklist item	Yes	No	Partial	NA
11	Hypothetical Survey response rates indicated: General Study response rate				
	WTP question response rate				
12	Actual survey (SME) response rates indicated: General Study response rate				
	WTP question response rate				
B. Criterion validity assessment elements					
13	Indicates whether assessment is between-sample or within-sample comparison.				
14	Indicates the duration between the hypothetical CV survey and actual (SME) survey.				
15	Describes the methods used to assess criterion validity.				
16	Indicates and justifies the values presented as comparators.				
C. **Study results**					
17	Reports the respondents' characteristics for both the hypothetical and actual (SME) survey				
18	Reports the summary estimates obtained and related of variability in the data for both the hypothetical and actual survey[43].				
19	If regression analysis is conducted, reports predictors and predictions of WTP for both the hypothetical and actual survey				
D. **Conclusions on criterion validity and declarations**					
20	Indicates the comparisons done for the assessment of criterion validity (e.g. mean hypothetical WP and mean values from actual survey).				
21	Decision on which validity will be confirmed provided (e.g. is this decision based on mean differences?).				
22	The decision on the criterion validity of CV-WTP from the study is clearly indicated.				
23	Declaration of conflict of interest (e.g. institutional affiliations, study funders).				

[43] Disaggregated estimates should be reported for all endpoints where multiple study attributes are evaluated (e.g. multiple study administration techniques, elicitation and analysis methods).

10.3 Limitations of the thesis

While this thesis has made significant contributions to the body of knowledge on the criterion validity of contingent valuation WTP studies, some limitations were encountered.

In conducting the systematic reviews presented in chapters 4 and 5, it is likely that some papers may have been missed. This omission is not considered as significant for the reviews of general validity as it is for the reviews on criterion validity. The review on general validity aimed at summarising the methods used to assess the different types of validity. As a result, given the breadth of methods already identified, it is likely that these have been exhausted. However, the variety in the terms used to describe criterion validity could have led to missed papers. As indicated in the respective chapters, robust searches including reference list, citation and author searches were more fruitful than the conventional database search. This process is believed to have been exhaustive in identifying all the relevant papers. Further, a citation alert set up on the key databases has not generated any additional papers that match the review criteria.

Further, the limited empirical assessments of the criterion validity of CV-WTP also narrowed the range of analyses that could be conducted to further investigate criterion validity assessments. One way that this was addressed was by using all the estimates reported in the studies, and thus a larger dataset. The challenges with this option relates to weighting of the studies in the analyses. To address this, during the analysis, comparisons were weighted by the studies. The inclusion of all the studies identified in the systematic review in chapter 5 into the quantitative synthesis in chapter 6 was not possible. This was because studies were either reported in ways that would not permit a synthesis, or data was missing. One way of addressing this might have been to contact the study authors for the missing information or clarifications where this was needed. However, as this situation affected only approximately one-fifth of the identified studies, it is believed that the results would not have changed significantly.

The use of a secondary data for the analyses presented in chapters 8 and 9 limited the range of analysis that could be conducted to the available variables. Primary data collection would have been more amenable to the exploration of a range of experimental protocol related to the conduct of CV-WTP and the effect of these on criterion validity assessments and conclusions. However, researchers agree that the use of secondary data is a cost-effective way of utilising all the available data on a subject (Cheng & Phillips 2014; Vartanian 2011). Further, the available dataset was relatively large and supported within a large clinical trial. This meant that robust checks were in place to ensure the quality of data. An empirical study at the time would not have been as robust and would have probably been much smaller.

The secondary dataset used for these analyses is also relatively old, having been collected in the year 2000. However, as the focus of the present analyses was primarily an investigation of the methods, with no interest in the actual WTP values for the mosquito nets, this dataset was considered appropriate for the analyses. Further, empirical assessments in the field have remained few. Even fewer are empirical assessments of hypothetical WTP employing multiple elicitation formats and estimates. This further limited the options for suitable datasets for this type of analyses. However, the choice of the dataset was guided by a systematic criteria, lending credibility to the process.

Further, the analyses presented were not guided by any CV-WTP models as these do not exist. It is therefore possible that the models were incorrectly specified, hence affecting the interpretations. However, the development of the models was informed by evidence from previous empirical studies and situated within the available theories, both economics and contingent valuation. Further, model diagnostics confirmed that the models were correctly specified. This further adds to the suggestion for guidelines for the conduct of and reporting of criterion validity CV studies, a draft of which have been provided in the previous section.

Finally, data from the simulated market experiment was not available. An analysis of these would have highlighted the closest prediction of actual value from the hypothetical WTP estimates, adding to the evidence on different elicitation techniques.

This thesis does not address the question of whether people do what they say they will do. However, the analysis presented sufficiently addresses the aims of the thesis. While acknowledging that the predictors of WTP are different across elicitation methods and bidding levels, the thesis assesses and establishes that these predictions are indeed different. The potential effect of these differences on criterion validity assessments is discussed.

10.4 Implications from the thesis for research

The findings reported in this thesis add to the knowledge and debates on the criterion validity of CV-WTP methods. Based on the work presented in this thesis, the following suggestions for future work will advance the knowledge in this subject.

A key finding from the systematic reviews of the evidence is that there is still limited empirical assessments of the criterion validity of CV-WTP in the health sector. Efforts to improve the method must involve further empirical work to establish an evidence base on the effect of different experimental protocol and methods on criterion validity. As has been suggested in earlier chapters, when multiple estimation techniques are used to elicit WTP, criterion validity assessments should be conducted and reported for all estimates. This permits interpretations of criterion validity at different elicitation levels and for the different methods. These are included in the suggested guidelines presented in section 10.2.

Empirical work on the assessment of criterion validity could include re-analysis of available secondary datasets. As discussed in chapter 7, the analysis of secondary datasets is a cost-effective way of fully utilising the available data. Such re-analyses could explore different methodological approaches, such as model specification. This will not only establish the evidence on the methods used in investigating criterion validity but may also lead to different conclusions on criterion validity in available empirical assessments.

Future systematic reviews and meta-analysis of empirical analyses of the criterion validity of CV-WTP methods will also need to employ more robust approaches such as reference list, author and citation searches. In addition, to increase the datasets for such systematic reviews, all reported estimates could be included in the synthesis, where these are comparable.

10.5 Recommendations for research

To further aid in the understanding and interpretation of criterion validity assessments, the following recommendations are offered.

1. Of great significance for the improvement of the method, some guidelines for the conduct and reporting of criterion validity assessments are necessary. This will make the identification of literature, interpretation and synthesis of results easier. Hopefully, this will translate into growth in criterion validity assessments and the synthesis of such studies. Section 10.2 presents some suggestions in this regard.
2. While market prices elicited through simulated market experiments are suggested as a valid criterion for the assessment of criterion validity, there is limited evidence to support this assertion. Future studies might explore this further to establish or strengthen the evidence on the use of simulated market experiments.
3. In addition, while critics argue that CV-WTP methods do not pass the criterion validity test, it is not clear in the literature what this pass mark should be. For instance, how close is close enough for criterion validity to hold? Indeed, should the hypothetical WTP estimates and actual values be the same? The reviewed literature presents validity conclusions based on seemingly random decisions, with validity confirmed at different ratios or mean differences. Empirical work to address this question is critical for the advancement of this method.
4. Finally, missing in the literature on criterion validity assessments also is a discussion on the appropriate duration of time between hypothetical WTP and actual surveys. Questions such as the ideal time difference between the two studies need to be answered. This is because when studies are held too close then respondents are likely to remember what they said in hypothetical surveys. This may be positive, in that they may be consistent in their response, or it may be negative because they would simply repeat what they said in the first survey without a careful consideration of their preferences, hence leading to incorrect welfare estimates. If conducted too far apart then respondents might have forgotten about the survey altogether, but the time may also provide the respondents with an opportunity to carefully interrogate their preferences considering their budgets. Further investigation of this will improve the method further.

10.6 Concluding remarks

The primary objective of this thesis was to critically evaluate the assessment of criterion validity in CV-WTP methods. Further, the thesis aimed to demonstrate that the current narrative that CV-WTP is not criterion valid may not be entirely correct; rather, the methods used to assess the criterion validity are wanting.

Evidently, a lot of work is needed to improve the design and execution of CV-WTP criterion validity assessments and the analysis of data thereof. A major challenge that was established in this thesis is the lack of guidelines on the conduct and reporting of criterion validity assessments. A recommendation from the work presented in this dissertation is the development of guidelines for the conduct and reporting of criterion validity assessments. Initial attempts to address this include the suggested guidelines for the reporting of criterion validity assessments outlined in section 10.2. The availability of guidelines will hopefully harmonise attributes such as validity definitions, the conduct of criterion validity assessments, the analysis and reporting of hypothetical WTP and actual values. Further, the analyses presented in this thesis points to the importance of careful design, analysis and reporting of hypothetical WTP surveys, in light of their significance in the criterion validity assessments.

References

AbouZahr, C. (1999). Disability adjusted life years (DALYs) and reproductive health: a critical analysis. *Reproductive Health Matters*, 7(14), pp.118–129.

Agresti, A. (1990). *Categorical Data Analysis*, New York: John Wiley & Sons, Inc.

Ahlheim, M., Ekasingh, B., Fror, O., Kitchaicharoen, J., Neef, A., Sangkapitux, C., & Sinphurmsukskul, N. (2010). Better than their reputation: enhancing the validity of contingent valuation mail survey results through citizen expert groups. Journal of *Environmental Planning and Management*, 53(2), pp.163–182.

Ajzen, I., Brown, T.C., & Carvajal, F. (2004). Explaining the Discrepancy between Intentions and Actions: The Case of Hypothetical Bias in Contingent Valuation. *Personality and Social Psychology Bulletin*, 30(9), pp.1108–1121.

Akter, S., Brouwer, R., Chowdhury, S., & Aziz, S. (2007). Testing Reliability and Construct Validity of In- kind WTP Responses in Contingent Valuation., *PREM Working Paper* 07-07.

Alberini, A. (1995). Optimal Designs for Discrete Choice Contingent Valuation Surveys: Single-Bound, Double-Bound, and Bivariate Models. *Journal of Environmental Economics and Management*, 28(3), pp.287–306.

Alberini, A., Boyle, K.J. & Welsh, M.P. (2003). Analysis of Contingent Valuation Data with Multiple Bids and Response Options Allowing Respondents to Express Uncertainty. J*ournal of Environmental Economics and Management*, 45(1), pp.40–62.

Alberini, A., Veronesi, M., & Cooper, J.J. (2005). Detecting Starting Point Bias in Dichotomous-Choice Contingent Valuation Surveys, *FEEM working paper* No. 119.05.

Aleme, A., Girma, E. & Fentahun, N. (2014). Willingness to pay for insecticide-treated nets in Berehet District, Amhara Region, Northern Ethiopia: implication of social marketing. *Ethiop J Health Sci*, 24(1), pp.75–84.

Alwin, D.F. (1992). Information Transmission in the Survey Interview : Number of Response Categories and the Reliability of Attitude Measurement *Socilogy and methodology*, Vol . 22, pp . 83-118.

Andersson, H. & Svensson, M. (2008). Cognitive ability and scale bias in the contingent valuation method. *Environmental and Resource Economics*, 39(4), pp.481–495.

Archer, K., & Lemeshow, S. (2006). *Goodness-of-fit test for a logistic regression model fitted using survey sample data*. Stata Journal, 6(1), pp.97–105.

Arndt, J. & Crane, E. (1975). Response Bias, Yea-Saying, and the Double Negative. *Journal of Marketing Research*, 12(2), pp.218–220.

Arrow, K., Solow, R., Portney, P.R., Leamer, E.E., Radner, R., & Schuman, H. (1993).

Report of the NOAA Panel on Contingent Valuation. *Federal Register*, 58(10), pp.4601–4614.

Bala, M. V., Mauskopf, J., & Wood, L.L. (1999). Willingness to Pay as a Measure of Health Benefits. PharmacoEconomics, 15(1), pp.9–18.

Balistreri, E., McClelland, 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. *Environmental and Resource Economics*, 18(3), pp.275–292.

Barron, A.C., Lee, T.L., Taylor, J., Moore, T., Passo, M.H., Graham, T.B., Griffin, T.A., Grom, A.A., Lovell, D.J., & Brunner, H.I. (2004). Feasibility and construct validity of the parent willingness-to-pay technique for children with juvenile idiopathic arthritis. *Arthritis Care and Research*, 51(6), pp.899–908.

Bateman, I., Munro, A., Rhodes, B., Starmer, C., & Sugden, R. (1997). A test of the theory of reference-dependent preferences. *Quarterly Journal of Economics*, 112, pp.479–505.

Bateman, I., Carson, R.T., Day, B., Hanemann, W.M., Hanley, N., Hett, T., Lee, M.J., Loomes, G., Mourato, S., Ozdemiroglu, E., & Pearce, D.W. (2002). *Economic Valuation with Stated Preference Techniques: A Manual*, Cheltenham: Edward Elgar.

Bateman, I. & Turner, R. (1993). The Contingent Valuation Method. In K. Turner, ed. *Sustainable Environmental Economics and Management: Principles and Practice*. London: Belhaven.

Baum, C.F. (2006). *An Introduction to Modern Econometrics Using Stata*, Stata Press.
Bayoumi, A.M. 2004. The measurement of contingent valuation for health economics. *PharmacoEconomics*, 22(11), pp.691–700.

Belotti, F., Deb, P., Manning, W.G., & Norton, E.C. (2015). twopm:two-part models. *Stata Journal*, 15(1), pp.3–20.

Bhatia, M. (2000). *Economic evaluation of malaria control interventions in Surat, India*. PhD Thesis, University of London.

Bhatia, M.R. (2005). From evidence to calibration for starting point bias: willingness to pay for treated mosquito nets in Gujarat, India. *Applied Economics*, 37(1), pp.1–7.

Bhatia, M.R. & Fox-Rushby, J.A. (2002). Willingness to pay for treated mosquito nets in Surat, India: the design and descriptive analysis of a household survey. *Health policy and planning*, 17(4), pp.402–411.

Bhatia, M.R. & Fox-Rushby, J.A. (2003). Validity of Willingness to Pay: hypothetical versus actual payment. *Applied Economics Letters*, 10(12), pp.737–740.

Bhattacherjee, A. (2012). *Social Science Research: Principles, Methods, and Practices* 2nd ed., CreateSpace Independent Publishing Platform.

Biadgilign, S., Reda, A.A. & Kedir, H. (2015). Determinants of willingness to pay for the retreatment of insecticide treated mosquito nets in rural area of eastern Ethiopia. *International journal for equity in health*, 14(1), p.99.

Bishop, R.C. & Heberlein, T.A. (1979). Measuring Values of Extramarket Goods: Are Indirect Measures Biased? *American Journal of Agricultural Economics*, 61(5), pp.926–930.

Bleichrodt, H. (1997). Health utility indices and equity considerations. *Health Economics*, 16(1), pp.65–91.

Blomquist, G.C. & Whitehead, J.C. (1998). Resource quality information and validity of willingness to pay in contingent valuation. *Resource and Energy Economics*, 20(2), pp.179–196.

Blumenschein Blomquist, GC., Johannesson, M., Horn, N., Freeman, P, K. (2008). Eliciting willingness to pay without bias: Evidence from a field experiment. *The Economic Journal*, (118), pp.114–137.

Blumenschein, K., Johannesson, M., Blomquist, G.C., Liljas, B., & O'Connor, R.M. (1998). Experimental Results on Expressed Certainty and Hypothetical Bias in Contingent Valuation. *Southern Economic Journal*, 65(1), pp.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. *Journal of health economics*, 20(3), pp.441–457.

BMJ. (2013). Consolidated Health Economic Evaluation Reporting Standards (CHEERS) statement. *BMJ*, 346(F1049).

Bobinac, A., van Exel, N.J., Rutten, F.F., & Brouwer, W.B. (2012). Get more, pay more? An elaborate test of construct validity of willingness to pay per QALY estimates obtained through contingent valuation. *Journal of Health Economics*, 31(1), pp.158–168.

Bohm, P. (1972). Estimating demand for public goods: An experiment. *European Economic Review*. 3(111).

Botelho, A. & Pinto, L.C. (2002). Hypothetical, real, and predicted real willingness to pay in open-ended surveys: experimental results. *Applied Economics Letters*, 9(15), pp.993–996.

Bowling, A. (2001). *Measuring disease. A review of disease-specific quality of life measurement scales.* Second edition., Maidenhead, GB: Open University Press.

Bowling, A. (2002). *Research methods in health: investigating health and health services*, Buckingham: Open University Press.

Boyle, K.J. & Özdemir, S. (2009). Convergent validity of attribute-based, choice questions in stated-preference studies. *Environmental and Resource Economics*, 42(2), pp.247–264.

Bradburn, M., Deeks, J. & Altman, D. (1998). Metan—An alternative meta-analysis command (PDF Download Available). *Stata Tech Bull*, 44(15).

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), pp.159–165.

Brazier Ratcliffe, J., Salomon, J., & Tsuchiya, A., J. (2007). *Measuring and Valuing Health Benefits for Economic Evaluation*, New York: Oxford University Press.

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. *Journal of Economic Behavior and Organization*, 43(1), pp.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 Economics*, 72(2), pp.152–166.

Bryan, S. & Jowett, S. (2010). Hypothetical versus real preferences: Results from an opportunistic field experiment. *Health economics*, 19, pp.1502–1509.

Bursac, Z., Gauss, C.H., Williams, D.K., & Hosmer, D.W. (2008). Purposeful selection of variables in logistic regression. *Source Code for Biology and Medicine*, 3(17).

Byrnes, B., Jones, C. & Goodman, S. (1999). Contingent Valuation and Real Economic Commitments: Evidence from Electric Utility Green Pricing Programmes. *Journal of Environmental Planning and Management*, 42(2), pp.149–166.

Calia, P. & Strazzera, E. (2001). A sample selection model for protest responses in contingent valuation analyses. *Statistical*, 61(3), pp.473–485.

Camacho-Cuena, E., Garcia-Gallego, A., Georgantzis, N & Sabater-Grande, G. (2004). An Experimental Validation of Hypothetical WTP for a Recyclable Product. *Environmental and Resource Economics*, 27(3), pp.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. *Applied Economics Letters*, 7(7), pp.447–450.

Carmines, E.G. & Zeller, R.A. (1979). *Reliability and validity assessment*, Beverly Hills; London: Sage Publications.

Carson, R.T. (2012). Contingent Valuation: A Practical Alternative when Prices Aren't Available. *Journal of Economic Perspectives*, 26(4), pp.27–42.

Carson, R.T. (2000). Contingent Valuation: A User's Guide. *Environmental Science &*

Technology, 34(8), pp.1413–1418.

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 Economics*, 72(1), pp.80–99.

Carson, R.T. & Louviere, J.J. (2011). A Common nomenclature for stated preference elicitation approaches. *Environmental and Resource Economics* 44(4), pp.539–559.

Chambers, C.M., Chambers, P.E. & Whitehead, J.C. (1998). Contingent Valuation of quasi-public good: validity, reliability, and application to valuing a historic site. *Public Finance Review*, 26(2), pp.137–154.

Champ, P.A., Bishop, R.C., Brown, T.C. & McCollum, D.W. (1997). Using Donation Mechanisms to Value Nonuse Benefits from Public Goods. *Journal of Environmental Economics and Management*, 33(2), pp.151–162.

Champ, P.A. & Bishop, R.C. (2001). Donation Payment Mechanisms and Contingent Valuation: An Empirical Study of Hypothetical Bias. *Environmental and Resource Economics*, 19(4), pp.383–402.

Champ, P.A., Boyle, K.J. & Brown, T.C. (eds). (2017). *A Primer on Nonmarket Valuation (The Economics of Non-Market Goods and Resources)* 2nd ed., Springer.

Chase, C., Sicuri, E., Sacoor, C., Nhalungo, D., Nhacolo, A., Alonso, P.L. & Menendez, C. (2009). Determinants of household demand for bed nets in a rural area of Southern Mozambique. *Malaria Journal*, 8(1), p.132.

Chatterjee, S., Hadi, A.S. & Price, B. (2000). *Regression analysis by example* 3rd ed., New York, US: John Wiley & Sons, Inc.

Chavas, J.-P., Bishop, R. & Segerson, K. (1986). Ex ante consumer welfare evaluation in cost-benefit analysis. *Journal of Environmental Economics and Management*, 13(3), pp.255–268.

Chen, G. & Tsurumi, H. (2010). Probit and Logit Model Selection. *Communications in Statistics - Theory and Methods*, 40, pp.159–175.

Cheng, H.G. & Phillips, M.R. (2014). Secondary analysis of existing data: opportunities and implementation. *Shanghai Archives of Psychiatry*, 26(6), pp.371–375.

Chern, W. & Kaneko, N. (2007). Some Problems in Estimating Willingness to Pay with Contingent Valuation Surveys: Case for Consumer Acceptance of Genetically Modified Food. In *Southern Agricultural Economic Association*.

Chien, Y.L., Huang, C.J. & Shaw, D. (2005). A general model of starting point bias in double-bounded dichotomous contingent valuation surveys. *Journal of Environmental Economics and Management*, 50(2), pp.362–377.

Clarke, P.M. (2002). Testing the convergent validity of the contingent valuation and travel cost methods in valuing the benefits of health care. *Health Economics*, 11(2), pp.117–127.

Cocheba, D. & Langford, W. (1978). Wildlife Valuation: The Collective Good Aspect of Hunting. *Land Economics*, 54(4), pp.490–504.

Couch, A. & Keniston, K. (1960). Yeasayers and naysayers: Agreeing response set as a personality variable. *Journal of Abnormal & Social Psychology*, 60(2), pp.151–74.

Cronbach, L.J. (1971). Educational measurement. In R. Thorndike, ed. *Educational Measurement*. Washington DC: American Council on Education.

Cronbach, L.J. & Meehl, P.E. (1955). Construct validity in psychological tests. *Psychological bulletin*, 52(4), pp.281–302.

Cummings, R.G., Harrison, G.W. & Rutstrom, E.E. (1995). Homegrown values and hypothetical surveys: Is the dichotomous choice approach incentive-compatible? *American Economic Association*, 85(1), pp.260–266.

Cummings, R., Brookshire, D. & Schulze, W. (1986). *Valuing Environmental Goods: A State of the Arts Assessment of the Contingent Valuation Method*, Totowa, NJ: Rowman and Allanheld.

Cummings, R.G., Elliott, S., Harrison, G.W. & Murphy, J. (1997). Are hypothetical referenda incentive compatible? *Journal of Political Economy*, 105(3), pp.609–621.

Dalecki, M., Ilvento, T. & Moore, D. (1988). The Effects of Multi-Wave Mailings on the External Validity of Mail Surveys. *Journal of Community Development Society*, 19(1).

Damschroder, L.J., Ubel, P.A., Riis, J. & Smith, D.M. (2007). An alternative approach for eliciting willingness-to-pay: A randomized Internet trial. *Judgment and Decision Making*, 2(2), pp.96–106.

Dash, A., Valecha, N., Anvikar, A.R. & Kumar, A. (2008). Malaria in India: Challenges and opportunities. *J. Biosci*, 33, pp.583–592.

Davis, R. (1964). The Value of Big Game Hunting in a Private Forest. In *Transactions of the 29th North American Wildlife and Natural Resources Conference*. Washington DC: Wildlife Management Institute.

Davis, R. (1963a). Recreation Planning as an Economic Problem. *Natural Resources Journal*, 3(2), pp.239–249.

Davis, R. (1963b). *The Value of Outdoor Recreation: An Economic Study of the Maine Woods*. Havard University.

Diamond, P.A. & Hausman, J.A. (1994). Contingent Valuation: Is Some Number better than No Number? *Economic Perspectives*, 8(4), pp.45–64.

Diener, A., O'Brien, B. & Gafni, A. (1998). Health care contingent valuation studies: a review and classification of the literature. *Health Economics*, 7(4), pp.313–326.

Dobson, A. & Barnett, A. (2008). *Introduction to Generalized Linear Models* 3rd ed., Boca Raton, FL: Chapman and Hall/CRC.

Donaldson, C., Jones, A.M., Mapp, T.J & Olson, J.A. (1998). Limited dependent variables in willingness to pay studies. *Applied Economics*, 30, pp.667–677.

Drèze, J. & Stern, N. (1987). The theory of cost-benefit analysis. In *Handbook of Public Economics*. Elsevier B.V., pp. 909–989.

Drummond, M.F., Stoddart, G. & Torrance, G. (1987). *Methods for the economic evaluation of health care programmes*, Oxford UK: Oxford University Press.

Edwards, R.T. (2001). Paradigms and research programmes: is it time to move from health care economics to health economics? *Health Economics*, 10(7), pp.635–49.

Edwards, R.T., Charles, J.M. & Lloyd-Williams, H. (2013). Public health economics: A systematic review of guidance for the economic evaluation of public health interventions and discussion of key methodological issues. *BMC Public Health*, 13(1001).

Fernando, R. (2011). Logit, Probit and Tobit Models for Categorical and Limited Dependent Variables. In *PLCS/RDC Statistics and Data Series*.

Fincham, J.E. (2008). Response Rates and Responsiveness for Surveys, Standards, and the Journal. *American Journal of Pharmaceutical Education*, 72(2), p.43.

Flachaire, E. & Hollard, G. (2007). Starting point bias and respondent uncertainty in dichotomous choice contingent valuation surveys. *Resource and Energy Economics*, 29(3), pp.183–194.

Foreit, J.R. & Foreit, K.G.F. (2003). The reliability and validity of willingness to pay surveys for reproductive health pricing decisions in developing countries. *Health policy*, 63(1), pp.37–47.

Fox-Rushby, J.A. (2002). *Disability adjusted life years (DALYs) for decision-making?: an overview of the literature*, London: OHE.

Fox, J.A., Shogren, J.F., Hayes, D.F. & Kliebenstein, J.B. (1998). CVM-X: Calibrating Contingent Values with Experimental Auction Markets. *American Journal of Agricultural Economics*, 80(3), pp.455–465.

Freeman III, A. (2003). *The Measurement of Environmental and Resource Values: Theory and Methods* Second edition, Routledge.

Frondel, M. & Vance, C. (2013). On Interaction Effects: The Case of Heckit and Two-Part Models, *EconStor Open Access Articles*, ZBW - Leibniz Information Centre for Economics, pages 22-38

Frykblom, P. (1997). Hypothetical Question Modes and Real Willingness to Pay. *Journal of Environmental Economics and Management*, 34(3), pp.275–287.

Gafni, A. (1997). Willingness to pay in the context of an economic evaluation of healthcare programs: theory and practice. *American Journal of Managed Care*, 3(Suppl), pp.S21-32.

Gafni, A. (1991). Willingness-to-Pay as a Measure of Benefits: Relevant Questions in the Context of Public Decisionmaking about Health Care Programs. *Medical Care*, 29(12), pp.1246–1252.

Gebresilassie, F. & Haile Mariam, D. (2000). Factors Influencing People's Willingness-to-buy Insecticide-treated Bednets in Arbaminch Zuria District, Southern Ethiopia. *J health popul nutr*, 29(3), pp.200–206.

Gelman, A. & Hill, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models (Analytical Methods for Social Research)*, Cambridge: Cambridge University Press.

Getzner, M. (2000). Hypothetical and Real Economic Commitments, and Social Status, in Valuing a Species Protection Programme. *Journal of Environmental Planning and Management*, 43(4), pp.541–559.

Gold, M.R., Siegel, J.E., Russel, L.B. & Weinstein, M.C. (1996). *Cost-Effectiveness in Health and Medicine*, Oxford University Press.

Graham, D. (1981). Cost-Benefit Analysis Under Uncertainty. *The American Economic Review*, 71(4), pp.712–725.

Gramlich, E. & Rubinfeld, D. (1983). Micro Estimates of Public Spending Demand Functions and Tests of the Tiebout and Median-Voter Hypotheses. *Journal of Political Economy*, 90(3), pp.536–560.

Greene, W.H. (2003). *Econometric Analysis* P. Education, ed., Prentice Hall.
Guion, R.M. 1974. Content Validity - The Source of My Discontent. *Training*, 1(1), pp.1–10.

Gujarati, D. (2003). *Basic Econometrics* 4th edition. London: McGraw-Hill.

Haab, T. & McConnell, K. (2002). *Valuing Environmental and Natural Resources The Econometrics of Non-market Valuation*, Cheltenham: Edward Elgar.

Haller, K.B. (1990). Research Instruments: Assessing Validity. *MCN, The American Journal of Maternal/Child Nursing*, 15(3), p.214.

Halstead, J., Luloff, A. & Stevens, T. (1992). Protest Bidders in Contingent Valuation. *Northeastern Journal of Agricultural and Resource Economics*, 21(2).

Hanemann, M. (1994). Valuing the Environment Through Contingent Valuation. *Economic Perspectives*, 8(4), pp.19–43.

Hanemann, M. (1984). Welfare evaluations in contingent valuation experiments with discrete responses. *American Journal of Agricultural Economics*, 66, pp.332–341.

Hanemann, W. (1991). Willingness to pay and willingness to accept: how much can they differ? *American Economic Review*, 81, pp.635–647.

Hardwick, P., Khan, B. & Langmead, J. (1986). *An introduction to modern economics*, London: Longman.

Harrell, F.E.J. (2016). *Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*, Springer International Publishing.

Harris, C., Drver, B. & McLaughlin, W. (1989). Improving the contingent valuation method: A psychological perspective. *Journal of Environmental Economics and Management*, 17, pp.213–219.

Harris, R., Bradburn, M.J., Deeks, J.J., Harbord, R.M., Altman, D.G. & Sterne, J.A.C. (2008). metan: fixed- and random-effects meta-analysis. *Stata Journal*, 8 (1), pp.3–28.

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

Hauber, A. (2009). Healthy-years equivalent: wounded but not yet dead. *Expert review of pharmacoeconomics & outcomes research*, 9(3), pp.265–9.

He, S., Florkowski, W. & Jordan, J. (2002). Irrational Responses in Contingent Valuation and Their Potential Impacts on Mean Stated Willingness to Pay. In *Exploring Diversity in the European Agri-Food System*.

Heberlein, T.A. & Bishop, R.C. (1986). Assessing the validity of contingent valuation: Three field experiements. , 56, pp.99–107.

Hergies Kling, C.L., Azevedo, C., J.A. (1999). *Linking Revealed and Stated Preferences to Test External Validity*, Iowa. USA: Iowa State University.

Herriges, J.A. & Shogren, J.F. (1996). Starting point bias in dichotomous choice valuation with follow-up questioning. *Journal of Environmental Economics and Management*, 30(1), pp.112–131.

Hicks, J. (1956). *A Revision of Demand Theory*, Clarendon Press.

Hicks, J. (1943). The Four Consumer's Surpluses. *The Review of Economic Studies*, 11(1), pp.31–41.

Hicks, J. (1941). The Rehabilitation of Consumers' Surplus. *The Review of Economic Studies*, 8(2), pp.108–116.

Hinkle, D.E., Wiersma, W. & Jurs, S. (2003). *Applied Statistics for the Behavioral Sciences, International Edition* 5th ed., Boston: Houghton Mifflin.

Hjelmgren, J., Berggren, F. & Andersson, F. (2001). Health Economic Guidelines - Similarities, Differencres and Some Implications. *Value in Health*, 4(3).

Hoevenagel, R. (1996). The Validity of the Contingent Valuation Method : Perfect and Regular Embedding. *Environment and Resource economics*, 7(1), pp.57–78.

Holmes, T. & Kramer, R. (1995). An Independent Sample Test of Yea-Saying and Starting Point Bias in Dichotomous-Choice Contingent Valuation. *Journal of Environmental Economics and Management*, 28, pp.121–132.

Hosmer, D. & Lemeshow, S. (2013). *Applied Logistic Regression*, New York: Wiley Blackwell.

India. (1958). *Manual of the Malaria Eradication Operation*, Malaria Institute of India.
India (1976). National Vector Borne Disease Control. *Ministry of Health and Family Welfare*.

Johannesson, M., Blomquist, G.C., Blumenschein, K., Johansson, P-o., Liljas, B. & O'Conor, R.M. (1999). Calibrating Hypothetical Willingness to Pay Responses. *Journal of Risk and Uncertainty*.

Johannesson, M. (1997). Some further experimental results on hypothetical versus real willingness to pay. *Applied Economics Letters*, 4(8), pp.535–536.

Johannesson, M., Liljas, B. & Johansson, P-o. (1998). An experimental comparison of dichotomous choice contingent valuation questions and real purchase decisions. *Applied Economics*, 30(5), pp.643–647.

Johannesson, M., Liljas, B. & O'Conor, R.M. (1997). Hypothetical versus real willingness to pay: some experimental results. *Applied Economics Letters*, 4(3), pp.149–151.

Johnston, R.J. (2006). Is hypothetical bias universal? Validating contingent valuation responses using a binding public referendum. *Journal of Environmental Economics and Management*, 52(1), pp.469–481.

Jones, A. (2007). *Applied Econometrics for HEalth Economists: A Practical Guide* 2nd ed., Chapmnan & Hall/ CRC.

Kane, M. (2001). Current concerns in Validity Theory. *Journal of Educational Measurement*, 38, pp.319–342.

Kaplan, R.M. & Saccuzzo, D.P. (1997). Psychological Testing (principles, applications and issues). *Journal of Chemical Information and Modeling*, 53(9), pp.1689–1699.

Kartman, B., Stålhammar, N.O. & Johannesson, M. (1996). Valuation of health changes with the contingent valuation method: a test of scope and question order effects. *Health economics*, 5(6), pp.531–541.

Kealy, M.J., Montgomery, M. & Dovidio, J.F. (1990). Reliability and predictive validity of contingent values: Does the nature of the good matter? *Journal of Environmental Economics and Management*, 19(3), pp.244–263.

Kelley, T. (1927). *Interpretation of educational measurements*, Oxord: World Book Co.

Kelly, M. & McDaid, D. 2005. *Economic appraisal of public health interventions*, NHS Development Agency.

Klose, T. (1999). The contingent valuation method in health care. *Health policy*, 47(2), pp.97–123.

Kumar, A., Valecha, N., Jain, T. & Dash, A.P. (2007). Burden of malaria in India: retrospective and prospective view. *Am J Trop Med Hyg*, 77, pp.69–78.

Kurth, A., Weaver, M, Lockhart, D. & Bielinski, L. (2004). The Benefit of Health Insurance Coverage of Contraceptives in a Population-Based Sample. *American Journal of Public Health*, 94, pp.1330–1332.

Kutluay, Y.M., Brouwer, R. & Tol, R.S.J. (2015). *Valuing malaria morbidity : Results from a global meta- analysis*, Department of Economics, University of Sussex.

Labelle, R. & Hurley, J. (1992). Implications of basing health-care resource allocations on cost-utility analysis in the presence of externalities. *Health Economics*, 11(3), pp.259–77.

Lal, S., Sonal, G. & Phukan, P. (2000). Status of malaria in India. *J Indian Acad Clin Med*, 5(1), pp.19–23.

Landy, F.J. (1986). Stamp collecting versus science: Validation as hypothesis testing. *American Psychologist*, 41(11), pp.1183–1192.

Lew, D.K. & Wallmo, K. (2011). External Tests of Scope and Embedding in Stated Preference Choice Experiments: An Application to Endangered Species Valuation. *Environmental and Resource Economics*, 48(1), pp.1–23.

Lienhoop, N. & Ansmann, T. (2011). Valuing water level changes in reservoirs using two stated preference approaches: An exploration of validity. *Ecological Economics*, 70(7), pp.1250–1258.

Liljas, B. & Blumenschein, K. (2000). On hypothetical bias and calibration on cost-benefit studies. *Health policy*, (52), pp.53–70.

List, J.A. (2001). Do Explicit Warnings Eliminate the Hypothetical Bias in Elicitation Procedures? Evidence from Field Auctions for Sportscards. *The American Economic Review*, 91(5), pp.1498–1507.

List, J.A. & Gallet, C.A. (2001). What Experimental Protocol Influence Disparities

Between Actual and Hypothetical Stated Values? *Environmental and Resource Economics*, 20(3), pp.241–254.

List, J.A. & Shogren, J.F. (1998). Calibration of the difference between actual and hypothetical valuations in a field experiment. *Journal of Economic Behavior and Organization*, 37(2), pp.193–205.

List, J.A. & Shogren, J.F. (2002). Calibration of Willingness-to-Accept. *Journal of Environmental Economics and Management*, 43(2), pp.219–233.

Little, J. & Berrens, R. (2003). Explaining disparities between actual and hypothetical stated values: Further investigation using meta-analysis. *Economics Bulletin*, 3(1).

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. *Journal of Agricultural and Applied Economics*, 41(3), p.697.

Loomis, J., Brown, T., Lucero, B. & Peterson, G. (1997). Evaluating the Validity of the Dichotomous Choice Question Format in Contingent Valuation. *Environmental and Resource Economics*, 10(2), pp.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 Economics*, 72(4), pp.450–461.

Loomis, J., Miller, J., Gonzalez-Caban, A. & Champ, J. (2006). Testing the Convergent Validity of Videotape Survey Administration and Phone Interviews in Contingent Valuation. *Society & Natural Resources*, 19(4), pp.367–375.

Macmillan, D.C., Smart, T.S. & Thorburn, A.P. (1999). A Field Experiment Involving Cash and Hypothetical Charitable Donations. *Environmental and Resource Economics*, 14(3), pp.399–412.

MacPhail, F. (1998). Moving Beyond Statistical Validity in Economics. *Social Indicators Research*, 45(1–3), pp.119–149.

Marjon, van-der P., Shiell, A., Au, F, Johnston, D. & Tough, S. (2008). Convergent validity between a discrete choice experiment and a direct, open-ended method: Comparison of preferred attribute levels and willingness to pay estimates. *Social Science and Medicine*, 67(12), pp.2043–2050.

Martín-Fernández, J., del Cura-González, M. I., Rodríguez-Martínez, G., Ariza-Cardiel, G., Zamora, J., Gómez-Gascón, T., Polentinos-Castro, E., Pérez-Rivas, F. J., Domínguez-Bidagor, J., Beamud-Lagos, M., Tello-Bernabé, M. E., Conde-López, J. F., Aguado-Arroyo, Ó., Sanz-Bayona, M. T. & Gil-Lacruz, A. I. (2013). Economic valuation of health care services in public health systems: a study about Willingness to Pay (WTP) for nursing consultations. *PloS one*, 8(4), e62840.

Mas-Colell, A., Whinston, M. & Green, J. (1995). *Microeconomic theory*, New York:

Oxford University Press.

Mataria, A., Donaldson, C., Luchini, S. & Moatti, J-P. (2004). A stated preference approach to assessing health care-quality improvements in Palestine: From theoretical validity to policy implications. *Journal of Health Economics*, 23(6), pp.1285–1311.

McClelland, G.H., Schulze, W.D., Lazo, J.K., Waldman, D., Doyle, J.K., Elliott, S.R., Irwin, J.R. & EPA Science advisory board. (1992). *Methods for Measuring Non-Use Values: A Contingent Valuation Study of Groundwater Cleanup*. EPA Working Paper No. EE-0013.

Mccollum, D.W. & Boyle, K.J. (2005). The effect of respondent experience/knowledge in the elicitation of contingent values: An investigation of convergent validity, procedural invariance and reliability. *Environmental and Resource Economics*, 30(1), pp.23–33.

McIntosh, E., Clarke, P.M., Frew, E.J. & Louviere (Eds). (2010). *Applied methods of cost-benefit analysis in health care*, Oxford University Press.

McIntosh, E., Donaldson, C. & Ryan, M. (1999). Recent advances in the methods of cost-benefit analysis in healthcare. Matching the art to the science. *PharmacoEconomics*, 15(4), pp.357–367.

Mehrez, A. & Gafni, A. (1991). The Healthy-years Equivalents: How to Measure Them Using the Standard Gamble Approach. *Medical Decision Making*, Vol 11(2), pp.140–146.

Messick, S. (1986). *The once and future issues of validity: Assessing the meaning and consequences of measurement*, New Jersey.

Messick, S. (1993). Validity. In R. Linn, ed. *Educational Measurement*. Washington DC: Oryx Press.

Mishan, E. (2016). *Elements of Cost-Benefit Analysis*, New York: Routledge.

Misra, S. (1999). *Indoor residual spray versus treated mosquito nets using deltamethrin to control malaria - A community randomized trial in rural Surat, India*. PhD Thesis. University of London.

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

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. & The PRISMA group. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Medicine*, 6(7).

Morton, L.M., Cahill, J. & Hartge, P. (2005). Reporting Participation in Epidemiologic Studies: A Survey of Practice. *American Journal of Epidemiology*, 163(3), pp.197–203.

Mozumder, P. & Berrens, R.P. (2007). Investigating hypothetical bias: induced-value tests of the referendum voting mechanism with uncertainty. *Applied Economics Letters*, 14(10), pp.705–709.

Mujinja, P.G.M. (2006). Exploring determinants of cunsumer preferences and willingness to pay for treated bednets before intervention in a poor rural Tanzania. *East African Journal of Public Health*. 3(1) pp. 17-23

Mukaka, M. (2012). A guide to appropriate use of Correlation coefficient in medical research. *Malawi Medical Journal*, 24(3).

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

Mulligan, P. (1978). *Willingness to Pay for Decreased Risk from Nuclear Plant Accidents*, INIS Working paper No. 43

Murphy, J.J., Stevens, T.H. & Weatherhead, D. (2002). *An Empirical Study of Hypothetical Bias in Voluntary Contribution Contingent Valuation: Does Cheap Talk Matter?* , World Congress of Environmental and Resource Economists.

Murphy, J.J., Allen, P.G., Stevens, T.H. & Weatherhead, D. (2005). A Meta-analysis of Hypothetical Bias in Stated Preference Valuation. *Environmental & Resource Economics*, 30(3), pp.313–325.

Murphy, J.J., Stevens, T.H. & Yadav, L. (2010). A Comparison of Induced Value and Home-Grown Value Experiments to Test for Hypothetical Bias in Contingent Valuation. *Environmental and Resource Economics*, 47(1), pp.111–123.

National Institute for Health and Care Excellence (NICE). (2013). Guide to the methods of technology appraisal.

Neill, H.R., Cummings, R.G., Ganderton, P.T., Harrison. & McGuckin, T. (1994). Hypothetical Surveys and Real Economic Commitments. *Land Economics*, 70(2), pp.145–154.

Neumann, P., Stone, P.W., Chapman, R.H. & Sandberg E.A. (2000). The quality of reporting in published cost-utility analyses, 1976-1997. *Ann Intern Med*, 132(964-72).

Nimdet, K., Chaiyakunapruk, N., Vichansavakul, K. & Ngorsuraches, S. (2015). A Systematic Review of Studies Eliciting Willingness-to-Pay per Quality-Adjusted Life Year: Does It Justify CE Threshold? *PLoS ONE,* 10(4).

Olsen, J.A. & Donaldson, C. (1998). Helicopters, Hearts and Hips: Using willingness to set priorities for public sector health care programmes. *Social Science & Medicine*, 46(1), pp.1–12.

Olsen, J.A. & Smith, R.D. (2001). Theory versus practice: a review of "willingness-to-pay" in health and health care. *Health Economics*, 10(1), pp.39–52.

Onwujekwe, O., Fox-Rushby, J. & Hanson, K. (2008). Construct Validity of the Bidding Game, Binary with Follow-up, and a Novel Structured Haggling Question Format in Determining Willingness to Pay for Insecticide-Treated Mosquito Nets. *Medical Decision Making*, 28(1), pp.90–101.

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? *Social Science and Medicine*, 60(3), pp.525–536.

Onwujekwe, O., Hanson, K. & Fox-Rushby, J. (2004). Inequalities in purchase of mosquito nets and willingness to pay for insecticide-treated nets in Nigeria: challenges for malaria control interventions. *Malaria journal*, 3(6).

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 Economics*, 13(5), pp.477–492.

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. *Social Science & Medicine*, 58(3), pp.525–537.

Onwujekwe, O., Chima, R., Shu, E., Nwagbo, D., Akpala, C. & Okonkwo, P. (2002). Altruistic willingness to pay in community-based sales of insecticide-treated nets exists in Nigeria. *Social Science and Medicine*, 54(4), pp.519–527.

Onwujekwe, O. & Nwagbo, D. (2002). Investigating starting point bias: a survey of willingness to pay for insecticide-treated nets. *Social Science and Medicine*, 55(12), pp.2121–21230.

Onwujekwe, O., Chima, R., Shu, E. & Nwagbo, D. (2001). Hypothetical and actual willingness to pay for insecticide-treated nets in five Nigerian communities. *Tropical Medicine & International Health*, 6(7), pp.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 Economics*, 10(2), pp.147–158.

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

Parker, R.M. (1990). Power, Control, and Validity in Research. *Journal of learning disabilities*, 23(10), pp.613–620.

Pattanayak, S., Sharma, V.P., Kalra, N.L. & Orlov, V.S. (1994). Malaria Paradigms in India and control strategies. *Indian Journal of malariology*, 31(4), pp.141–99.

Payne, K., McAllister, M. & Davies, L. (2013). Valuing the economic benefits of complex interventions: when maximising health is not sufficient. *Health Economics*, 22(3), pp.258–71.

Philips, Z., Whynes, D.K. & Avis, M. (2006). Testing the construct validity of willingness to pay valuations using objective information about risk and health benefit. *Health Economics*, 15(2), pp.195–204.

Portney, P.R. (1994). The Contingent Valuation Debate: Why Economists Should Care. *Economic Perspectives*, 8(4), pp.3–17.

Ramke, J., Palagyi, A., du Toit, R. & Brian, G. (2009). Stated and Actual Willingness to Pay for Spectacles in Timor-Leste. *Ophthalmic epidemiology*, 16(4), pp.224–230.

Reaves, D.W., Kramer, R.A. & Holmes, T.P. (1999). Does Question Format Matter? Valuing an Endangered Species. *Environmental and Resource Economics*, 14, pp.365–383.

Rennie, D. & Luft, H. (2000). Pharmacoeconomic analyses. *JAMA*, 283, pp.2158–60.
Ried, W. (1998). QALYs versus HYEs - what's right and what's wrong. A review of the controversy. *Health Economics*, 17(5), pp.607–25.

Roberts, K., Thompson, M. & Pawlyk, P. (1985). Contingent Valuation of Recreational Diving at Petroleum Rigs, Gulf of Mexico. *Transactions of the American Fisheries Society*, 114(2).

Rolfe, J. & Dyack, B. (2010). Testing for convergent validity between travel cost and contingent valuation estimates of recreation values in the Coorong, Australia. *Australian Journal of Agricultural and Resource Economics*, 54(4), pp.583–599.

Rosen, A., Greenberg, D., Stone, P.W., Olchanski, N.V. & Neumann, P.J. (2005). Quality of abstracts of papers reporting original cost-effectiveness analyses. *Medical Decision Making*, 25(4), pp.424–8.

Ryan, M., Mentzakis, E., Jareinpituk, S. Cairns, J. (2016). External Validity of Contingent Valuation: Comparing Hypothetical and Actual Payments. *Health Economics*, 26(11).

Samuelson, P. (1947). *Foundations of Economic Analysis*, Harvard University Press.

Sanders, G.D., Neumann, P.J. & Basu, A. (2016). Recommendations for Conduct, Methodological Practices, and Reporting of Cost-effectiveness Analyses. Second Panel on Cost-Effectiveness in Health and Medicine. *JAMA*, 316(10), pp.1093–1103.

Seip,K & Strand, J. (1992). Willingness to pay for environmental goods in Norway: A contingent valuation study with real payment. J*ournal of Environmental and Resource Economics*, 2(1) pp.91-106.

Severens, J., de Boo, T., van Roosmalen, M., Verweiji, P.E., vad der Wilt, G.J. (2000). Validity of willingness-to-pay for nondecisional diagnostic information. *HEPAC*, 1(9), pp.9–13.

Sharma, G. (1986). *Malaria and its control in India*, Delhi.

Sharma, R. (1995). *Malaria Action Programme based on expert committee report*, Delhi, India.

Sharp S. (1998). Meta-analysis regression. *Stata Tech Bull*, (42), pp.16–24.

Silberberg, E. (1978). *The Structure of Economics: A Mathematical Analysis* Third., New York: McGraw-Hill.

Slothuus, U., Larsen, M.L. & Junker, P. (2002). The contingent ranking method - A feasible and valid method when eliciting preferences for health care? *Social Science and Medicine*, 54(10), pp.1601–1609.

Smith, R.D. (2001). The relative sensitivity of willingness-to-pay and time-trade-off to changes in health status : an empirical investigation. *Health economics*, 10(6), pp.487–497.

Smith, V.K., Desvousges, W.H. & Freeman III, A. (1985). *Valuing Changes in Harzadous Waste Risks: A Contingent Valuation Approach*, N.C.

Song, L., Langfelder, P. & Horvath, S. (2013). Random generalized linear model: a highly accurate and interpretable ensemble predictor. *BMC Bioinformatics*, 14(5).

Soto Montes De Oca, G. & Bateman, I.J. (2006). Scope sensitivity in households' willingness to pay for maintained and improved water supplies in a developing world urban area: Investigating the influence of baseline supply quality and income distribution upon stated preferences in Mexico City. *Water Resources Research*, 42(7), pp.1–15.

Spencer Swallow, S.K., Miller, C.J., M.A. (1998). Valuing Water Quality Monitoring: A Contingent Valuation Experiment Involving Hypothetical and Real Payments. *Agricultural and Resource Economics Review*.

Streiner, D.L. (1989). *Health measurement scales: a practical guide to their development and use*, Oxford UK: Oxford University Press.

Streiner, D.L., Norman, G.R. & Cairney, J. (2008). *Health measurement scales: a practical guide to their development and use*, Oxford UK: Oxford University Press.

Sugden, R . (1999). Alternatives to the neoclassical theory of choice. In I. Bateman & K. Willis, eds. *Valuing Environmental Preferences: Theory and Practice of the Contingent Valuation Method in the US, EU, and Developing Countries*. Oxford: Oxford University Press, pp. 152–180.

Sugden, R. & Williams, A. (1978). *The Principles of Cost-Benefit Analysis*, Oxford University Press.

Taye, B. (2002). Willingness to Pay for Insecticide- Impregnated Bed Nets : The Case of Selected Rural Kebeles in Ilu Woreda of Western Shoa Zone. *Ethiopian Journal of Economics*, XI(1), pp.1–32.

Telser, H., Becker, K. & Zweifel, P. (2008). Validity and Reliability of Willingness-to-Pay Estimates from Two Overlapping Discrete- Choice Experiments. *Economic Affairs*, (0412).

Thorndike, E.L. (1913). *An introduction to the theory of mental and social measurements*, Columbia: Teachers College Columbia University.

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 Economics*, 22(4), pp.428–450.

Tussupova, K., Berndtsson, R., Bramryd, T. & Beisenova, R. (2015). Investigating willingness to pay to improve water supply services: Application of contingent valuation method. *Water (Switzerland)*, 7(6), pp.3024–3039.

UN. (2017). *World Population Prospects: The 2017 Revision, Key Findings and Advance Tables.* Working Paper No. ESA/P/WP/248.

Varian, H. (2014). *Intermediate microeconomics: a modern approach*, New York: W.W. Norton & Company.

Varian, H.R. (2006). Revealed Preference. *Samuelsonian economics and the twenty-first century*, (January 2005), pp.1–23.

Vartanian, T.P. (2011). *Secondary Data Analysis*, New York NY.

Veisten, K., Hoen, H.F., Navrud, S. & Strand, J. (2004). Scope insensitivity in contingent valuation of complex environmental amenities. *Journal of Environmental Management*, 73(4), pp.317–331.

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. *Applied Economics*, 38(7), pp.735–756.

Vernazza, C.R., Wildman, J.R., Steele, J.G., Whitworth, J.M., Walls, A.W.G., Perry, R., Mathews, R., Hahn, P. & Donaldson, C. (2015). Factors affecting patient valuations of caries prevention: Using and validating the willingness to pay method. *J Dent*, 43(8), pp.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. *Environmental and Resource Economics*, 49(2), pp.199–215.

Vossler, C.A.. Poe, G.L., Welsh, P. & Ethier, R.G. (2004). Bid Design Effects in Multiple Bounded Discrete Choice Contingent Valuation. *Environmental & Resource Economics*, 29(4), pp.401–418.

Vossler, C.A., Kerkvliet, J., Polasky, S. & Gainutdinova, O. (2003). Externally validating contingent valuation: an open-space survey and referendum in Corvallis, Oregon. *Journal of Economic Behavior and Organization*, 51(2), pp.261–277.

Vossler, C.A., Ethier, R.G., Poe, G.L. & Welsh, M.P. (2003). Payment Certainty in Discrete Choice Contingent Valuation Responses: Results from a Field Validity Test. *Southern Economic Journal*, 69(4), pp.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. *Journal of Environmental Economics and Management*, 45(3), pp.631–649.

Vossler, C.A. & Watson, S.B. (2013). Understanding the consequences of consequentiality: Testing the validity of stated preferences in the field. *Journal of Economic Behavior & Organization*, 86, p.137.

Wakker, P. (1996). A Criticism of Healthy-years Equivalents. *Medical Decision Making,* 16(3), pp.207–214.

Walsh, R., Miller, N. & Gilliam, L. (1983). Congestion and Willingness to Pay for Expansion of Skiing Capacity. *Land Economics*, 59(2), pp.195–210.

Weatherly, H., Drummond, M., Claxton, K., Cookson, R., Ferguson, B., Godfrey, C.A., RIce, N., Sculpher, M.J. & Sowden, A. (2009). Methods for assessing the cost-effectiveness of public health interventions: key challenges and recommendations. *Health Policy*, 93(2–3), pp.85–92.

Weinstein, M.C. & Stason, W.B. (1977). Foundations of cost-effectiveness analysis for health and medical practices. *New England Journal of Medicine*, 296(13), pp.716–21.

Welle, P. (1985). *Potential Economic Impacts of Acid Rain in Minnesota: The Minnesota Acid Rain Survey*, Minnesota.

Welsh, M.P. & Poe, G.L. (1998). Elicitation Effects in Contingent Valuation: Comparisons to a Multiple Bounded Discrete Choice Approach. *Journal of Environmental Economics and Management*, 36, pp.170–185.

Whitehead, J.C., Huangm J-C., Blomquist, G. & Ready, R. (1998). Construct validity of dichotomous and polychotomous choice contingent valuation questions. *Environmental and Resource Economics*, 11(1), pp.107–116.

WHO. (2016). WHO malaria terminology. *WHO*. Available from http://apps.who.int/iris/bitstream/handle/10665/208815/WHO_HTM_GMP_2016.6_eng.pdf;sequence=1. [Accessed on December 12, 2016].

WHO. (2008). The Global Burden of Disease: 2004 Update. Disease incidence, prevalence and disability. *Health statistics and information systems*. pp.28–37. Available from http://www.who.int/healthinfo/global_burden_disease/2004_report_update/en/. [Assessed on June 16, 2016].

WHO. (2012). Metrics: Disability-Adjusted Life Year (DALY). Health statistics and health information systems. Available from http://www.who.int/healthinfo/global_burden_disease/metrics_daly/en/. [Assessed on June 16, 2016].

Willis, K. & Powe, N. (1998). Contingent Valuation and Real Economic Commitments: A Private Good Experiment. *Journal of Environmental Planning and Management*, 41(5), pp.611–619.

Willingness-to-Pay [online]. (2016). York; York Health Economics Consortium; 2016. Available from https://www.yhec.co.uk/glossary/willingness-to-pay/ [Assessed on December 19, 2016].

Worthing, C.R., Walker, S.B. & British Crop Protection Council. (1983). *The Pesticide Manual: A World Compedium* 7th edition. The British Crop Protection Council.Croydon.

Yasunaga, H., Ide, H., Imamura, T. & Ohe, K. (2006). Contingent valuation for health care services. Review of domestic studies and outline of foreign investigations. *[Nippon kōshū eisei zasshi] Japanese journal of public health*, 53(11), pp.818–830.

Yeung, R.Y.T. & Smith, R.D. (2005). Can we use contingent valuation to assess the demand for childhood immunisation in developing countries?: a systematic review of the literature. *Applied health economics and health policy,* 4(3), pp. 165-173

Yue, A. (2010). Validity. In A. J. Mills Durepos, G. and Wiebe, E. (eds). *Encyclopaedia of Case Study Research*. Thousand Oaks, CA: Sage Publications.

Appendices

Appendix 1: Economic evaluation techniques

Random utility / Discrete choice models: These models describe the choices respondents make, given a set of alternatives which include other competing options or choices (Train, 2002; Small & Rosen, 1981).

Travel Cost Method (TCM): The travel cost method (TCM) is used to estimate the consumptive value of environmental attributes of goods. The method recognises that users pay an implicit price by giving up time and money to take trips to these areas for recreation. The cost of a visit to a site is the out-of-pocket costs of travel including any site admission fees, opportunity cost of travel time, and the opportunity cost of time on site. The travel cost itself is not the value of the resource – but this information is used to derive a demand curve to then estimate lower bound values for the resource (Bateman et al. 2002, Michell & Carson, 1989).

Hedonic Pricing: Hedonic pricing methods seek to exploit possible relationships between demands for private goods and their associated bundle of characteristics, including environmental characteristics. It uses, for example, information on people's job and location choices to estimate marginal willingness to pay for resource allocation changes (Bateman et al. 2002, Michell & Carson, 1989).

Averting Behavior: Averting behaviour models simulate consumer behaviour and rely on the existence of an activity that substitutes for the services provided by a resource e.g. environmental good. The averting behaviour method infers values from defensive, mitigating, or averting expenditures, i.e., those actions taken to prevent or counteract the adverse effects of environmental degradation (Bateman et al. 2002, Michell & Carson, 1989).

Market Prices: Market prices can be used to obtain individual’s willingness to pay (WTP) directly for goods and services. For the accurate assessment of this, demand (and not just market prices), needs to be assessed (Bateman et al. 2002, Michell & Carson, 1989).

Choice Modelling: People who are expected to experience the benefits or costs are asked a series questions about their preferences for alternative future goods or services. Each of the questions posed represents a choice set. Further, each choice set represents the outcome from one of the alternatives under valuation which are described as attributes (Bateman et al. 2002, Michell & Carson, 1989).

Contingent Valuation: Survey methods are used to elicit hypothetical estimates of value. Willingness to pay or willingness to accept techniques are employed (Bateman et al. 2002, Michell & Carson, 1989).

Appendix 2: WTP/WTA elicitation formats

1. **Open ended question (OE)**: In this, respondents are asked to state their maximum WTP for the amenity to be valued (McIntosh et al. 2010). A single question is asked and actual WTP estimates obtained. This format is often used in mail surveys and interviews.

2. **Bidding game format (BG):** In this, the question is designed so that it resembles an auction as the respondent enters a bargaining process with the interviewer. The respondent is presented with a first bid and depending on whether they accept or reject the bid it is either raised or lowered till eventually the respondent's maximum WTP is reached (McIntosh et al. 2010; Bateman et al. 2002). This format is best administered through interviews.

3. **Dichotomous choice methods (DC):** In these formats a discrete indicator of WTP is obtained (Bateman et al. 2002; McIntosh et al. 2010; Carson 2000). Variations of this format include:

    - Closed – ended single question in which the bid value is presented leading to a Yes or No answer. This is the simplest of these methods and can be administered through mail surveys or interviews.

    - Closed ended with a follow up question. This is an extension of the closed ended method aimed at obtaining additional information from each respondent by adding a follow up question to the closed-ended single question format above. The design of the question is a form of bidding truncated at two bids. This format is best administered through interviews.

    - Double bounded dichotomous choice. This involves an iterated series of questions in which respondents are asked additional questions if they would pay a higher or lower amount, depending on their responses to previous questions. This format is best administered through interviews.

4. **Payment Card (PC):** In this, respondents are presented with a range of values to choose from. A typical design presents respondents with a series of bid amounts in a vertical list from the lowest bid to the highest bids in increments (McIntosh et al. 2010). This format could be administered through mail or interviews.

5. **Single binary discrete choice format**: In the simplest format of this question, the respondent is presented with two alternatives leading to a Yes or No response (McIntosh et al. 2010; Bateman et al. 2002; Carson 2000). Owing to the simplicity of its design, this format can be administered through mails or interviews.

6. **Structured Haggling (SH):** The SH method resembles the BG and SBDC + OE but allows more steps the BG and SBDC +OE to mimic the haggling process if needed, so that respondents that are willing to pay are coaxed to state the highest possible amount they can pay (Onwujekwe et al. 2008). This format is best administered through interviews.

Appendix 3: Validity definitions

Face Validity. The most basic form of validity, face validity is concerned with how well a measure represents an intuitive and common-sense understanding of a phenomenon or how well a test or the questions on a test appear to measure the desired qualities of a particular construct (Bowling 2002; Carmines & Zeller 1979; Yue 2010; Streiner et al. 2008). Face validity considers whether the items of each domain are sensible and appropriate for a specific population (Brazier et al. 2007). The determination of face validity of a measure is based on its examination by persons with expertise in the health condition or intervention being measured such as the patients experiencing a disease for which a measurement scale is undergoing validation. The test for face validity depends purely on the judgment of the observer and is therefore highly subjective. Qualitative methods mainly focus group discussions have been used in face validity tests. Face validation of measurement scales is not common in the literature.

Content validity: As with face validity, content validity is a technical description of the judgement that a measure or scale looks reasonable (Brazier et.al. 2007). This refers to the extent to which the content of the measure adequately covers or represents all the relevant or important concepts of the subject being studied and is sufficiently sensitive to changes (Brazier et al. 2007; Bowling 2002; MacPhail 1998; Cronbach & Meehl 1955; Streiner 1989). Unlike face validity which is subjectively judged by common sense, a determination of content validity involves a review of the test content by subject-matter experts. The experts objectively judge whether a test or scale adequately reflects the content that is being measured (Cronbach & Meehl 1955). Assessment of content validity involves the development of a model by experts in the field which details the full domain of content that is relevant to the particular measurement, sampling specific relevant words from the collection and collating these in a testable manner (Carmines & Zeller 1979; Streiner et al. 2008).

Construct validity. This is also referred to in the literature as measurement validity and theoretical validity. Construct validity is defined as a measure of the degree to which a test assesses the underlying theoretical or hypothetical constructs it is supposed to measure (Cronbach & Meehl 1955; Streiner et al. 2008). Construct validity focuses on whether the measurement of one concept is related to that of other concepts in logical, predictable ways and is determined by examining whether a measure agrees with other measures or indicators of the subject being measured e.g. health (Brazier et al. 2007). This type of validity focusses on the testing of hypotheses (Streiner et al. 2008).

Construct validity is classified into divergent (or discriminant) and convergent validity demonstrated by the correlation of the measures used within constructs. For divergent validity to hold, measures of constructs that theoretically should not be related to each other should be in fact, observed to not be related to each other. For a measure to be regarded as demonstrating convergent validity, it should be possible to discriminate between dissimilar constructs (Parker 1990; Streiner et al. 2008; Haller 1990).

Criterion validity: Criterion validity refers to the correlation between the research instrument being studied and an established measure of the same concept – an externally-defined "gold-standard" (Carmines & Zeller 1979; Streiner et al. 2008; Haller 1990). This is also referred to in the literature as external validity. In contingent valuation studies, in the absence of a "gold standard", actual willingness to pay values are used as the criterion with which hypothetical willingness to pay values are assessed. Criterion validity is classified into predictive and concurrent validity. Predictive validity is a measure of the extent to which a future level of a variable can be predicted from a current measurement (MacPhail 1998; Carmines & Zeller 1979; Cronbach & Meehl 1955; Streiner et al. 2008). Concurrent validity is used where a criterion exists in the present. This measure looks for correlation with other tests. It is a measure of a tests' ability to distinguish between groups that it should theoretically be able to distinguish between and is assessed by correlating the measure and the criterion at the same time or in close proximity to one another (Carmines & Zeller 1979; Cronbach & Meehl 1955). The main difference between predictive and concurrent validity is the time when the test is administered with predictive validity used where a criterion is not available in the present.

Appendix 4: Sample search strategy from Medline database

1 (Will* and PAY).m_titl. (938)
2 (Will* and Accept).m_titl. (89)
3 "Contingent Val*".m_titl. (147)
4 "Hypothetical market*".m_titl. (0)
5 "Hypothetical Valu*".m_titl. (1)
6 "Stated Preference*".m_titl. (65)
7 "Stated Valu*".m_titl. (2)
8 1 or 2 or 3 or 4 or 5 or 6 or 7 (1185)
9 "Valid*".m_titl. (65916)
10 "Construct Val*".m_titl. (1241)
11 "Criterion Val*".m_titl. (192)
12 "Content Val*".m_titl. (329)
13 "Face Val*".m_titl. (112)
14 "Discriminant Val*".m_titl. (252)
15 "Convergent Val*".m_titl. (208)
16 "Theoretical Val*".m_titl. (46)
17 (Sensitivity and Scope).m_titl. (9)
18 "Psychometr*".m_titl. (7068)
19 "Psychological test*".m_titl. (775)
20 9 or 10 or 11 or 12 or 13 or 14 or 15 or 16 or 17 or 18 or 19 (73362)
21 8 and 20 (13)
22 (Will* and PAY).m_titl. (938)
23 (Will* and Accept).m_titl. (89)
24 "Contingent Val*".m_titl. (147)
25 "Hypothetical market*".m_titl. (0)
26 "Hypothetical Valu*".m_titl. (1)
27 "Stated Preference*".m_titl. (65)
28 "Stated Valu*".m_titl. (2)
29 22 or 23 or 24 or 25 or 26 or 27 or 28 (1185)
30 "Valid*".m_titl. (65916)
31 "Construct Val*".m_titl. (1241)
32 "Criterion Val*".m_titl. (192)
33 "Content Val*".m_titl. (329)
34 "Face Val*".m_titl. (112)
35 "Discriminant Val*".m_titl. (252)
36 "Convergent Val*".m_titl. (208)
37 "Theoretical Val*".m_titl. (46)
38 (Sensitivity and Scope).m_titl. (9)
39 "Psychometr*".m_titl. (7068)
40 "Psychological test*".m_titl. (775)
41 30 or 31 or 32 or 33 or 34 or 35 or 36 or 37 or 38 or 39 or 40 (73362)
42 29 and 41 (13)
43 (Will* and PAY).m_titl. (938)
44 (Will* and Accept).m_titl. (89)
45 "Contingent Val*".m_titl. (147)
46 "Hypothetical market*".m_titl. (0)
47 "Hypothetical Valu*".m_titl. (1)
48 "Stated Preference*".m_titl. (65)
49 "Stated Valu*".m_titl. (2)
50 43 or 44 or 45 or 46 or 47 or 48 or 49 (1185)
51 "Valid*".m_titl. (65916)
52 "Construct Val*".m_titl. (1241)
53 "Criterion Val*".m_titl. (192)
54 "Content Val*".m_titl. (329)
55 "Face Val*".m_titl. (112)
56 "Discriminant Val*".m_titl. (252)
57 "Convergent Val*".m_titl. (208)
58 "Theoretical Val*".m_titl. (46)
59 "Psychometr*".m_titl. (7068)
60 "Psychological test*".m_titl. (775)
61 51 or 52 or 53 or 54 or 55 or 56 or 57 or 58 or 59 or 60 (73362)
62 50 and 61 (93)

Appendix 5: Data extraction form

Variable	Definition and example where necessary
Study Aim and Country of study	Specific to study
Study type and Sample size	Qualitative/ Quantitative and Cross section/ Other
Study Perspective	Ex Ante or Ex Post; WTP or WTA
Intervention characteristics	Type (Environment/ Health/ Other) and Product/ Services (Specified)
Respondent Characteristics	Survey respondent (Household head/ Other), Sex, Experience with intervention
Study administration	Face to face; internet based; telephone interviews; postal mail surveys
Validity assessment method	Study hypothesis
Validity tests	Regression analyses
WTP Elicitation format	Dichotomous choice; open ended; payment card; binary game; binary with follow up
Payment vehicle	Out of pocket payment; additional tax; Voluntary contributions
Payment frequency	Weekly, monthly, yearly
Payment duration	One off, n number of months/ years
Sensitivity/ Bias / analysis	Hypothetical bias, starting point bias etc.
WTP/ WTA estimates	Mean or median
Results on validity tests	Specific validity test results
Authors conclusion on validity	Authors conclusion regarding validity tested
Reviewer comments	Reviewers comments on article

Appendix 6: List of included studies

1. Philips Z., Whynes, D.K & Avis, M. (2006) 'Testing the construct validity of willingness to pay valuations using objective information about risk and health benefit', *Health Econ* 02; 15(2):195-204.
2. Barron, A.C., Lee, T., Taylor, J., Moore, T., Passo, M.H., Graham, T.B., Griffin, T.A., Grom, A.A., Lovell, D.J. & Brunner, H.I. (2004) 'Feasibility and construct validity of the parent willingness-to-pay technique for children with juvenile idiopathic arthritis', *Arthritis Rheum (Arthritis Care Res)* 12/15; 51(6):899-908.
3. Onwujekwe, O., Fox-Rushby, J. & Hanson, K. (2008) 'Construct validity of the bidding game, binary with follow-up, and a novel structured haggling question format in determining willingness to pay for insecticide-treated mosquito nets', *Med Decis Making.* 28(1):90-101.
4. 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 Economics*. 74(4), pp. 450-461.
5. 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', *Journal of Environmental Economics and Management*. 45(3) pp. 631-649.
6. McCollum, D.W. & Boyle, K.J. (2005), 'The Effect of Respondent experience/ Knowledge in the Elicitation of Contingent Values: An Investigation of Convergent Validity, Procedural Invariance and Reliability', *Environmental & Resource Economics.* 30(1) pp. 23-33.
7. Rolfe, J. & Dyack, B. (2010). 'Testing for convergent validity between travel cost and contingent valuation estimates of recreation values in the Coorong, Australia', *Australian Journal of Agricultural and Resource Economics*. 54(4 pp. 583.
8. Marjon van der Pol, Shiell, A., Au, F., Johnston, D. & Tough, S. (2008). 'Convergent validity between a discrete choice experiment and a direct, open-ended method: Comparison of preferred attribute levels and willingness to pay estimates', *Social science & medicine*. 67(12). pp. 2043-2050.
9. Boyle, K.J. & Özdemir, S. (2009). 'Convergent Validity of Attribute-Based, Choice Questions in Stated-Preference Studies', *Environmental and Resource Economics*. 42(2). pp. 247-264.
10. Vossler, C.A., Ethier, R.G., Poe, G.L. & Welsh, M.P. (2003). 'Payment Certainty in Discrete Choice Contingent Valuation Responses: Results from a Field Validity Test', *Southern Economic Journal*. 69(4) pp. 886-902.
11. Champ, J., Miller, J., Gonzalez-Caban, A. & Loomis, J. (2006). 'Testing the Convergent Validity of Videotape Survey Administration and Phone Interviews in Contingent Valuation', *Society & Natural Resources*. 19(4). 4, pp. 367-375.
12. Sangkapitux, C., Neef, A., Kitchaicharoen, J., Sinphurmsukskul, N., Frör, O., Ekasingh, B. & Ahlheim, M. (2010). 'Better than their reputation: enhancing the validity of contingent valuation mail survey results through citizen expert

groups', *Journal of Environmental Planning and Management*. 53(2). pp. 163-182.

13. Clarke, P.M. (2002). 'Testing the convergent validity of the contingent valuation and travel cost methods in valuing the benefits of health care'. *Health Economics*. 11(2). pp. 117-127.
14. Chambers, C.M., Chambers, P.E. & Whitehead, J.C., 'Contingent Valuation of Quasi-Public Goods: Validity, Reliability, and Application to Valuing a Historic Site'. Working Papers 9614, East Carolina University, Department of Economics.
15. Whitehead, J.C., Huang, J.C., Blomquist, G.C. & Ready, R. C. (1998). 'Construct Validity of Dichotomous and Polychotomous Choice Contingent Valuation Questions'. *Environmental and Resources Economics*. 11(1). pp. 107-116
16. Severens, J.L., de Boo, Th.D., van Roosmalen, P., Verweij, E. & van der Wilt., G.J. (2000). 'Validity of Willingness-to-Pay for Nondecisional Diagnostic Information'. *Health Economics*. 1(1). pp. 9-13
17. Hoevenagel, R. (1996). 'The validity of the contingent valuation method: Perfect and regular embedding', *Environmental & Resource Economics*. 7(1). pp. 57-78.
18. Smith, R.D. (2001). 'The relative sensitivity of willingness-to-pay and time-trade-off to changes in health status: an empirical investigation', *Health Economics*. 10(6). pp. 487-497.
19. Hergies, J.A., Kling, C.L. & Azevedo, C. (1999). 'Linking Revealed and Stated Preferences to Test External Validity', Iowa State University, Working Paper 99-WP-222
20. Kartman, B., Stålhammar, N.O. & Johannesson, M. (1996). 'Valuation of health changes with the contingent valuation method: a test of scope and question order effects', *Health Economics*. 5(6). pp. 531-541.
21. Lienhoop, N. & Ansmann, T. (2011). 'Valuing water level changes in reservoirs using two stated preference approaches: An exploration of validity', *Ecological Economics*. 70(7). pp. 1250-1258.
22. Vossler, C.A. & Watson, S.B. (2013). 'Understanding the consequences of consequentiality: Testing the validity of stated preferences in the field', *Journal of Economic Behavior & Organization*. vol. 86, pp. 137.
23. Lew, D.K. & Wallmo, K. (2011), 'External Tests of Scope and Embedding in Stated Preference Choice Experiments: An Application to Endangered Species Valuation', *Environmental and Resource Economics*. 48(1). pp. 1-23.
24. Soto, M de Oca. G. & Bateman, I.J. (2006). 'Scope sensitivity in households' willingness to pay for maintained and improved water supplies in a developing world urban area: Investigating the influence of baseline supply quality and income distribution upon stated preferences in Mexico City', *Water Resources Research*. 42(7). .

25. Blomquist, G.C. & Whitehead, J.C. (1998). ‘Resource quality information and validity of willingness to pay in contingent valuation’, *Resource and Energy Economics*. 20(2). pp. 179-196.
26. Camacho-Cuena, E., García-Gallego, A., Georgantzís, N. & Sabater-Grande, G. (2004). ‘An Experimental Validation of Hypothetical WTP for a Recyclable Product’, *Environmental and Resource Economics*. 27(3). pp. 313-335.
27. Mataria, A., Donaldson, C., Luchini, S. & Moatti, J. (2004). ‘A stated preference approach to assessing health care-quality improvements in Palestine: from theoretical validity to policy implications’, *Journal of health economics*. 23(6). pp. 1285-1311.
28. Telser, H., Becker, K., & Zweifel P. (2009). ‘Validity and reliability of willingness-to-pay estimates: evidence from two overlapping discrete-choice experiments’, *Pharmaceutical Medicine*. 23(1). pp. 49.
29. Veisten, K., Fredrik Hoen, H., Navrud, S. & Strand, J. (2004). ‘Scope insensitivity in contingent valuation of complex environmental amenities’, *Journal of environmental management*. 73(4). pp. 317-331.
30. Foreit, J.R. & Foreit, K.G.F. (2003). ‘The reliability and validity of willingness to pay surveys for reproductive health pricing decisions in developing countries’, *Health policy*. 63(1). pp. 37-47.
31. 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’, *Journal of Agricultural and Applied Economics*. 41(3). pp. 697.
32. Loomis, J., Brown, T., Lucero, B. & Peterson, G. (1997). ‘Evaluating the Validity of the Dichotomous Choice Question Format in Contingent Valuation’, *Environmental and Resource Economics* 10(2). pp. 109-123.
33. Johnston, R.J. (2006). ‘Is hypothetical bias universal? Validating contingent valuation responses using a binding public referendum’, *Journal of Environmental Economics and Management*. 52(1). pp. 469-481
34. Fox-Rushby, J.A. & Bhatia, M.R. (2003). ‘Validity of Willingness to Pay: hypothetical versus actual payment’, *Applied Economics Letters*,10(12). pp. 737-740.
35. Vernazza, C.R., Wildman, J.R., Steele, J.G., Whitworth, J.M., Walls, A.W.G., Perry, R., Mathews, R., Hahn, P., Donaldson, C. (2015). ‘Factors affecting patient valuations of caries prevention: Using and validating the willingness to pay method’, *Journal of Dentistry*. 43(8). pp. 981-988
36. Muller, L. & Ruffieux, B. (2011). ‘Do price-tags influence consumers’ willingness to pay? On the external validity of using auctions for measuring value’, *Experimental Economics*. 14(2). pp. 181-202.
37. Vossler, C.A., Kerkvliet, J., Polasky, S. & Gainutdinova, O. (2003). ‘Externally validating contingent valuation: an open-space survey and referendum in Corvallis, Oregon’, *Journal of Economic Behavior and Organization*. 51(2). pp. 261-277.

Appendix 7: Sample search strategy from Medline database

1 (Will* and PAY).m_titl. (7600)

2 "Contingent Val*".m_titl. (538)

3 "Hypothetical market*".m_titl. (10)

4 "Hypothetical Valu*".m_titl. (43)

5 "Stated Preference*".m_titl. (444)

6 "Stated Valu*".m_titl. (61)

7 1 or 2 or 3 or 4 or 5 or 6 (8179)

8 "Valid*".m_titl. (526110)

9 "Construct Val*".m_titl. (13927)

10 "Criterion Val*".m_titl. (2983)

11 "External Val*".m_titl. (6490)

12 8 or 9 or 10 or 11 (526489)

19 7 and 12 (453)

20 Limit to reviews or synthesis (13)

Appendix 8: List of the reviews on criterion validity

1. 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 Economics*, **72**, 80-99.

2. Harrison, G.W. & Rutström, E.E. (2008). Experimental Evidence on the Existence of Hypothetical Bias in Value Elicitation Methods, in Smith, V. L., ed, *Handbook of Results in Experimental Economics*. New York: Elsevier Science.

3. Liljas, B & Blumenschein K. 2000. On hypothetical bias and calibration in cost-benefit studies. *Health Policy* 52:53–70.

4. List, J. & Gallet, C. 2001. What Experimental Protocol Influences Disparities between Actual and Hypothetical Stated Values? *Environmental and Resource Economics* 20: 241-254.

5. Little, J. & Berrens, R. (2004) "Explaining Disparities between Actual and Hypothetical Stated Values: Further Investigation Using Meta−Analysis." *Economics Bulletin*. 3(6) pp. 1−13.

6. Murphy, J.J., Allen, P.G., Stevens, T.H. & Weatherhead, D. (2005). “A Meta-Analysis of Hypothetical Bias in Stated Preference Valuation,” *Environmental and Resource Economics* 30:313-325*.*

Appendix 9: Detailed results on the review of reviews

From the total of 13 papers identified through all the processes, only six met the inclusion criteria and data from these was extracted. The reviews are summarised in chronological order as the reviewer sought to demonstrate the progression in the criterion validity assessment and agreements during the time period. For each review a brief introduction summarising the review objectives and methods is provided. This is followed by a discussion of the results, key findings and a conclusion. A list of the reviewed papers is provided in Appendix 3.10.

1. Carson et al. (1996)

Carson (1996) published the first quantitative synthesis of primary studies comparing hypothetical and actual values (herein referred to as an examination of external validity). The objective of this review was to summarise available information to provide readers with the broadest possible overview of how CV estimates for quasi-public goods correspond with estimates obtained from revealed preference techniques. The review identified 83 studies with a total of 616 comparisons of hypothetical and actual values. Stated preferences (CV) were compared against (actual) values obtained for the same quasi-public good using any revealed preference technique. The review considered only WTP estimates obtained from interviews with consumers and not governments or institutions. Identified studies spanned across the period 1966 to 1994 and were classified broadly into: recreational goods where activities such as sport fishing, hunting and camping were valued; environmental amenities where studies valued changes in goods such as air and water quality and health, for studies valuing small reductions in environmental or work-related health risks. Revealed preference techniques were coded into five broad categories: (1) Single site travel cost models (TC1); (2) Multiple travel cost models (TC2); (3) Hedonic pricing (HP); (4) Averting behaviour models (Avert) and; (5) Simulated or actual markets ACTUAL) for the good.

The authors included all available comparisons in the studies and those that were easily inferred. Multiple estimates from a single study were provided where: (1) a single study valued multiple goods such as where respondents were interviewed at several recreational fishing locations and TC and CV estimates made at each location; (2) where different levels of goods were valued; or (3) when a study used different analytical assumptions in making the CV/RP estimates.

The data was summarised in three ways: (i) using the complete dataset where each CV/revealed preference (RP) ratio was treated as an observation; (ii) using a 'trimmed dataset' where the remaining data after trimming off the smallest 5 percent and largest 5 percent of the CV/RP ratios was used; (iii) using a 'weighted sample', with the mean CV/RP ratio from each study's observation. Estimates included in each of these were the mean, standard error of the mean, the maximum, minimum and median (50th percentile) observations, a wide range of other percentiles of the sample distribution and the sample size. Four comparisons of CV and RP estimates were conducted: (a) Using the CV as the numerator; (b) Using the RP as the numerator; (c) Directly testing whether the quantity (CV-RP) is different from zero and; (d) Traditional vote-counting analysis which ignores the magnitude of difference by assigning a value of 1 for comparisons where CV is greater than 1 and 0 otherwise.

Results (a) using the CV as the numerator: (1) Complete dataset: mean CV/RP ratio: 0.890; CI (0.813-0.960); median: 0.747; (2) Trimmed sample: mean CV/RP ratio: 0.744; CI (0.736-0.811); median: 0.747; (3) Weighted sample: mean CV/RP ratio: 0.922; CI (0.811-1.034); median: 0.936. CV/RP ratios suggest that CV estimates are on average lower than their RP counterparts. Results (b) using the RP as the numerator: (1) Complete dataset: mean RP/CV ratio: 5.567; CI (4.189-7.153); median: 1.388; (2) Trimmed sample: mean RP/CV ratio: 2.626; CI (2.351-2.902); (3) Weighted sample: mean RP/CV ratio: 3.542; CI (2.029-5.057); median: 1.416. RP/CV ratios suggest that RP estimates are on average considerably larger than their CV counterparts. Results (c) directly testing whether the quantity (CV-RP) is different from zero rejects the null hypothesis in favour of the alternative that the difference is negative with t-statistics of: -7.31 (complete dataset); -6.19 (trimmed dataset); and -2.58 (weighted dataset). Results (d) using the traditional vote-counting analysis the null hypothesis that the vote-count is equal to zero can be rejected using a sign test in favour of the alternative that the average is less than zero with z-statistics of: 17.13 (complete dataset); 15.57 (trimmed dataset); and 5.44 (weighted dataset).

The authors conducted four different analyses:

1. Four (4) bivariate regressions. The authors did not provide a justification for the choice of the regressors in any of these. The bivariate regressions are conducted using: - (a) *RP technique used*. With the coefficients defined relative to the TC1 category (which was omitted), results suggest that the CV estimates run about 20% lower than the TC1 counterparts; 30% lower than their TC2 counterparts, a little less than 40% lower than their HP counterparts, about 20% lower than their AVERT counterparts and are on average indistinguishable from their ACTUAL counterparts. (b) *Broad class of goods valued*: With recreation goods (REC) used as the reference category, HEALTH goods may have CV/RP ratios closer to 1.0 relative to the other two categories of goods. (c) *Study publication status*: With non-published studies used as the reference category, results suggest that CV/RP ratios from published studies are closer to 1.0 compared with those from studies that are not published. (d) *Time periods*. Dummy categories used in this analysis were: (i) studies published (or unpublished, dated) prior to 1984; (ii) studies published between 1984 and 1989; and (iii) studies published after 1989. The results suggest that the CV/RP ratios do not exhibit any statistically significant difference between the three time-periods.

2. A meta-analysis was not conducted owing to incomplete reporting of the necessary details. However, from the bivariate analysis, the authors identified that single-site TC1 produced higher CV/RP ratios on average than did multiple-site models (TC2). This is largely because many TC1 models did not include any value for travel time while most TC2 models make some allowances for travel time.

3. Correlation coefficients were estimated using two methods. For each sample, the correlation coefficients are provided: Complete sample: - Pearson coefficient: 0.83; Spearman’s coefficient: 0.78; Trimmed sample: - Pearson coefficient: 0.91; Spearman’s coefficient: 0.88; Weighted sample: - Pearson coefficient: 0.98; Spearman’s coefficient: 0.92. In all three datasets, both correlation coefficients are significantly different from zero (p<0.001) which suggests that if the RP estimates are systematically varying with the nature of the good being valued, then so are the CV estimates.

4. Regressing the RP estimate on the CV estimate, the coefficient on the CV estimate ranges from 0.9 to 1.4 and is always highly significant, depending on the sample used. The intercept term is always positive and tends to be reasonably large

and quite significant for treatments where the coefficient on the CV estimate is near or below 1.0. The best fitting regression model was found by taking the average RP and CV estimates from the 83 studies as the observations when the averaging is performed using the trimmed dataset rather than the complete dataset. The high R^2 of 0.98 generated from the analysis suggests that, after eliminating a fraction of the between studies variance by trimming off the overall smallest and largest 5% of the CV/RP rations and eliminating all within study variance by averaging, the CV and RP estimates are very closely linked.

The authors of this review conclude that, based on the CV RP comparisons summarised, arbitrarily discounting CV estimates by a factor of two or more, as proposed (by the NOAA panel), appears to be unwarranted. This is because CV/RP ratios of >2.0 comprise only 5% of the complete sample and only 3% of the weighted sample in this review. Applying a discounted factor of 2.0 or greater to the CV estimates used in the analysis would result in "adjusted" CV estimates that, in almost all cases, diverge from the estimates obtained from observable behaviour, rather than converge.

2. ***Harrison and Rutström, 2008***

Glenn and Rutström (2008) conducted a narrative review of the basic experimental results that support the conclusion that hypothetical bias exists. This review was conducted in 1999 but was not published until 2008. Recognising that several published experimental results already confirmed that hypothetical bias exits, the authors' clustered thirty-five (35) studies based on the type of goods and value elicitation mechanisms. This formed an early assessment on the effect of type of goods (public versus private) and two WTP elicitation mechanisms (open ended (OE) and dichotomous choice (DC)) on hypothetical bias. The study inclusion criteria and the estimates included in the summary are not described in the paper, but the reviewed studies were published between the years 1972 and 1998.

The review studies were grouped into: (1) CV literature and tests with private goods; (2) CV literature and tests with public goods; (3) Open-ended elicitation in the lab; (4) Dichotomous choice elicitation in the lab and (5) Social elicitation in the lab. Some of these studies are discussed further in the review. Further, the authors summarised the responses that have been proposed to mitigate hypothetical bias: instrument and statistical calibration.

(1) CV studies with private goods. In their study, Dickie, Fisher and Gerking (1987) obtained values for a pint of strawberries by using CVM, and also by actually selling the strawberries to households. The authors conclude that they could not reject the null hypothesis of structurally identical demand equations estimated from actual sales and CVM data. Harrison and Ruestrom however find that using the same dataset, the hypothetical demand curve can overstate the quantity demanded depending on the price used. The hypothetical bias calculated from the raw data is approximately 58% with the review authors concluding that there is unequivocal support in the study for the view that hypothetical and actual questions generate the same demand schedules. In their studies, Bishop and Heberlein (1979) and Bishop, Heberlein, and Kealy (1983), found evidence of hypothetical bias in CVM estimates for subjects' WTA for returning their goose hunting permits and WTA values based upon actual cash payments. In a re-evaluation of the results by Bishop and Heberlein (1979), Hanemann (1984) further demonstrated the extreme sensitivity of the study's conclusions to alternative statistical assumptions. Bishop and Heberlein (1986) found that on average, hypothetical WTP values exceeded real ones.

(2) CV studies with public goods. Using a closed-circuit broadcast of a new Swedish TV program as the valuation commodity, Bohm (1972) obtained CVM and actual values in an experiment in which he used six elicitation procedures to evaluate strategic bias. In each case except one, the TV program was made available and subjects in each group allowed to see it, if aggregate WTP equalled or exceeded a known total cost. No formal theories were provided to generate the hypothesis for the procedures used in the experiments. Based on a parametric analysis, Bohm concluded that bids were virtually identical for all institutions. However, the samples are not normally distributed and thus the parametric test was not ideal. By conducting a non-parametric test on the same data, Cummings and Harrison (1994) found evidence of hypothetical bias in all but one procedure and concluded that not all of the hypothetical bias can be explained by strategic bias. Kealy, Montgomery and Dovidio [1990] examined the predictive validity of CVM values for actual cash payment for both a private good (a chocolate bar) [72

respondents] and a public good (a de-acidification program for lakes in the Adirondack region) [107 respondents]. The authors report significant hypothetical bias of 30% for both the private and public goods. Using a dichotomous choice question, Seip and Strand (1992) elicited WTP for contributions towards nature conservation from a sample of 101 Norwegians. Based on the results, the authors conclude that hypothetical bias was 2,017%. In their study, Duffield and Patterson (1992) used mail surveys to obtain three sets of values for a fund to be established for the purpose of leasing water rights to be used for the preservation of in-stream flows in a set of Montana rivers. In one set, respondents were asked to make an actual tax-deductible contribution while a second set was asked whether they would make the contribution if contacted within a month to make the tax-deductible contribution. In the third set, a group of the respondents in the first two sets was revisited four years later to confirm whether had acted as they had claimed they would. A hypothetical bias of 35% was observed in the first two sets. For the third set, 91% of the respondents had acted as they said they would while among those who claimed they would connect, 29% never did.

Based on the review of CVM studies conducted for public and private goods, Glenn and Rutström concluded that there was evidence for hypothetical bias in valuation tasks regardless of the type of good. However, the authors acknowledged that the results were difficult to interpret and also sensitive to variations in the different experimental designs and field conditions under which the studies were conducted. In attempts to explain the difficulties in interpretation, the authors reviewed studies that conducted laboratory experiments using open-ended, dichotomous choice and social elicitation methods.

(3) Open-ended elicitation in the Lab. Using vickery auctions, Neil et al (1994) elicited values from subjects for a small oil painting by an unknown Navajo artist, or a reprint of a medieval map. The purpose of this experiment was to see how much of the hypothetical bias was due to the hypothetical nature of the economic commitment, as revealed by the difference between hypothetical vickery auctions (HVA) and real vickery auctions (RVA) and how much would be due to the absence of the structured institution in a CVM as revealed by the difference between HVA and the CVM. Three values were elicited: An unstructured CVM where the subject was asked to state the maximum amount they would be willing to pay for the painting;

hypothetical and real vickery auctions where the former was identical to the latter with instructions minimally changed to reflect the hypothetical nature of the transaction. Hypothetical bias of over 2,400% was established between the HVA and RVA and 2,600% between the CVM and the RVA for the same good. Hypothetical bias for the painting was 290%. The authors concluded that the lack of real commitment in either of the two hypothetical institutions was the culprit for hypothetical bias. In an experiment using induced and home-grown values for an insurance product, McClelland, Schulze and Coursey (1993) asked respondents to bid for insurance policies to avoid some low-probability bad outcome that was generated by the authors according to a specified probability distribution. The authors found that hypothetical bias changed with the probability function such that at low probabilities hypothetical WTP was about twice that of the actual bids but that this bias is reduced, then eliminated and in some cases reversed, as the probability of a loss increases. The highest bias found was 120% for inexperienced responses to the lowest risk event. Further, for the two highest risk events the inexperienced responses show that real bids exceed hypothetical bids by about 25%.

(4) Dichotomous choice elicitation in the lab. Cummings, Harrison and Rutström (1995) randomly assigned subjects to one of two treatments with the only difference being the use of hypothetical or real language in the instructions. The valuation goods were an electric juicer, chocolates and a calculator. Hypothetical respondents responded much more positively than real subjects leading the experimenters to reject incentive compatibility. Hypothetical bias was found as 163% (juicers), 873% (chocolates) and 163% (calculator). In their study, Johannesson, Liljas and Johansson (1998), make some wording changes to the earlier study by Cummings, Harrison and Rutstrom (1995). They followed up all hypothetical "yes" responses by asking subjects to state if they were "fairly sure" or "absolutely sure" they would buy the good. By taking only the latter responses as indicating a "yes", the authors conclude that hypothetical bias disappears. Using a DC design, Smith and Mansfield (1998) asked subjects who had just participated in an interview if they would be willing to participate in another one in the future for compensation amounts ranging from $5 up to $50. Based on their results the authors conclude that there is strong evidence of the absence of hypothetical bias. In his study, Frykblom (1997) uses both OE and DC questions

to elicit values for a private good: an environmental atlas that retails for SEK 200. The hypothetical bias based on the OE survey is 50% and 56% based on the DC and both are significant.

(5) Social elicitation in the lab. Cummings, Elliott, Harrison and Murphy [1997] undertook majority rule experiments for an actual public good. After earning some income, in addition to their show-up fee, subjects were asked to vote on a proposition that would have each of them contribute a specified amount towards this public good and if the majority said “yes”, all had to pay. The authors found hypothetical bias of 67% and this was significant.

Harrison and Rutström (1999) also reviewed papers in which attempts had been made to calibrate the hypothetical values, to control for it. The two main calibration methods discussed were ‘instrument’ and ‘statistical’ calibration methods. ‘Instrument’ calibration involves ex-ante attempts to choose suitable words during the design of the questionnaire which are aimed at encouraging the respondent to reveal their true values. This calibration method also involves choosing different word formats in a laboratory setting to test for the reduction of hypothetical bias. Commonly used scripts include certainty questions (how certain are you that you would pay $X). ’Statistical’ calibrations are ex-post processes which involve statistical analyses to determine whether observed hypothetical bias is systematic or predictable, and adjusting for this during the analysis phase.

Three papers that have attempted to elaborate on the subject of statistical calibrations were reviewed. Blackburn, Harrison and Rutström (1994) offer the analogy of a watch that is always 10 minutes slow to introduce the idea of a statistical bias function. The authors argue that hypothetical responses can still be informative if the bias between the two is systematic and predictable. They further define a “known bias function” as one that is a systematic statistical function of the socio-economic characteristics of the sample. The authors show that one can use the bias function estimated from one instance to calibrate the hypothetical responses in another instance and that the calibrated responses statistically match those observed in a paired real elicitation procedure. However, this test was limited to private goods only. Harrison et al. (1998) undertook five surveys with each designed to provide information that would allow for the calibration of a field public good. In one survey, a deliverable private good was used, and a vickery auction with real payment required. The next survey varied this by allowing for public provision of a private good with real payment required. Both free riding and hypothetical bias were measured by comparing the results of these two surveys. Arguing that the propensity to engage in hypothetical bias would be independent of the propensity to engage in free-riding bias, the results from the initial two surveys were then used to adjust the results in the latter two surveys used to elicit hypothetical bias with the end result being a statistical measure of the propensity of subjects to engage in both types of bias. This measure was used to adjust hypothetical responses to a survey asking for valuation of a non-deliverable public good. The results also demonstrate the potential complementarity between lab and field experiments.

Fox et al. (1998) discuss and apply a calibration approach to hypothetical survey values that uses experiments to ascertain possible hypothetical bias. Using health risk reduction in a food product (irradiated food product versus the raw food product) as the commodity, hypothetical valuations were obtained from 174 pork-eating respondents. Further lab experiments were conducted with respondents who agreed to participate from the initial experiment. After five (5) rounds of bidding, subjects were given information on the difference between the two products and allowed to bid for five (5) more rounds. One of the rounds was chosen at random and the transactions effected. The results suggest that calibration factors of roughly 2/3 with comparisons of hypothetical survey values and the round 2 auction values. The auction values in the final round are generally higher than those in round 2, so the calibration factors are higher as well – between 60% and 83%.

In their conclusion, the authors of this review confirm the presence of hypothetical bias. They also identify the variety of elicitation formats, subject pools, and the type of good (private or public) as potential drivers of hypothetical bias. The authors also conclude that based on the available experimental evidence, the treatment of simple “yes” and definitely “yes” responses impact on conclusions on hypothetical bias, and that the extent of hypothetical bias varies from setting to setting. However, the authors assert that the sample sizes and designs employed as at the time of their review were far too slight for one to draw any broad conclusions from them. The authors argue that “it is particularly inappropriate to try to claim that hypothetical bias is any more of a problem in open ended formats as compared to closed ended formats, or that referendum formats are “less inaccurate” than DC formats. Lastly, the authors note that some calibration methods could reduce hypothetical bias.

3. ***Liljas, Bengt and Blumenschein Karen, 2000***

In contributing to the growing body of evidence, Liljas, Bengt and Blumenschein Kares (2000) reviewed hypothetical bias in economic experiments performed to compare real and hypothetical WTP and further sought to discuss what could be done to eliminate or reduce this difference. The authors do not indicate the criteria used to identify the included studies or the time periods. The authors do not also discuss the estimates that have been compared in the analysis. A total of nineteen (19) studies published between 1972 and 1999 are included in the review. In summarising the evidence, studies were clustered according to the elicitation methods: (1) Open ended WTP questions; (2) Auctions and (3) Dichotomous choice WTP questions.

1. Open-ended WTP questions. Three (3) studies were reviewed (Bohm, 1972; Seip and Strand, 1992 and Navrud, 1992). The authors noted that while studies using the OE method were few as at the time of the review, results showed evidence of hypothetical bias. In his study, Bohm (1972) ran five experiments[44] in which he compared different real WTP elicitation methods. In another experiment, he compared the real WTP methods to a hypothetical WTP elicitation method. Bohm found that the five real WTP methods did not differ significantly, but that one of them differed significantly when compared to the hypothetical method. Seip and Strand (1992) attempted to look at possible differences in real and hypothetical WTP for a membership in an environmentalist association. Their results showed poor correspondence between real and hypothetical WTP. In his study, Navrud (1992) investigated the WTP for preserving endangered species using an

[44] Elicitation Procedures [Bohm (1972)]

Procedure 1: The subject paid according to his stated WTP.

Procedure 11: The subject paid some fraction (less than 1) of his stated WTP, with the fraction determined equally for all in the group such that total costs are just covered.

Procedure 111: Subjects did not know the specific payment scheme at the time of their bid, but did know that it was a lottery with equal probability attached to the payment schemes of procedures 1, 11, IV and V.

Procedure IV: Each subject paid a fixed amount (SEK 5).

Procedure V: The subject paid nothing.

elicitation technique similar to a payment card. The results of his analysis showed a better correspondence than those of Seip and Strand.

2. Experiments with auctions. Five (5) studies investigated the validity of open-ended CV questions in vickery auctions (Neill et al. (1994); Loomis et al (1996); Blumenschein et al (1997); Johannensson (1997) and Johannesson et al. (1997). All the studies used private goods and the participants were randomized into hypothetical and real groups. The first four studies found large and significant differences in real and hypothetical WTP whereas the fifth study found a close correspondence.
3. Experiments with dichotomous choice WTP questions. Eleven (11) studies which elicited WTP using DC questions were reviewed. In valuing goose permits, Bishop and Heberlein (1990) found a poor correspondence between real and hypothetical WTP. In further analysis of the same study, the authors found a close correspondence between the WTP assessed by both OE and DC questions and the real WTP but a poor correspondence with WTA. In their study, Dickie et al (1987) found that the values did not differ significantly. On the contrary, other studies [Cummings et al (1995), Johannesson et al. (1998), Blumenschein et al. (1998), Champ et al. (1997), Nape et al. (1995), Brown et al. (1996), Cummings et al. (1997) and Bjornstad et al. (1997)] found that hypothetical values overestimated real values.

The review authors further sought to summarize the evidence on studies that had calibrated responses to correct discrepancies between hypothetical and real WTP. Studies were grouped into those that calibrated hypothetical WTP responses with a control group and those that did not include a control group. In calibrations of hypothetical WTP responses without a control group, the individuals were given several choices on how to answer the WTP questions in order to be able to find out the degree of uncertainty about their answers to the WTP questions. In investigating the presence of an ambivalent region where individuals are ambivalent about answering yes or no to the WTP question, the authors used a polychotomous question which included multiple yes responses (such as maybe yes, certainly yes). By treating all yes responses as “yes”, Ready et al (1995) found that where ambivalence exists, the likely effect would be to decrease the regular DC estimate of the mean WTP. To handle yea-saying, Eckerlund et al. (1995) and Kartman et al

(1996), using a polychotomous question used a conservative approach to take the yea-saying behaviour into account with the effect that the mean WTP was significantly reduced. Blamley et al (1999) used a dissonance minimizing approach which involves giving the respondents other non-monetary options to choose from in attempts to handle yea-saying. The authors found that the method produced higher price sensitivity and a lower mean WTP as compared to both the DC and the polychotomous format.

In calibrating hypothetical WTP responses with a control group, Johannesson et al (1998) tested a more conservative interpretation of the DC approach where only “definitely sure” yes responses were counted as “yes” responses. In this study, the number of “definitely sure” responses were found to significantly underestimate the number of real yes responses and thus provide a lower bound for the real WTP. Using the same calibration method, Blumenschein et al. (1998) found that there was no longer a significant difference between the real and hypothetical WTP responses. In a different study, Champ et al. (1997) assessed the certainty of hypothetical donation responses on a scale of 1-10 (ranging from “very uncertain” to “very certain”) and found that hypothetical donations significantly exceeded real donations but there was no significant difference if only subjects that were very certain of their yes-responses were counted as real yes-responses.

In a different calibration, Fox et al (1998) used a technique they referred to as CVM-X where a sub-sample of the respondents in the hypothetical component participates in the real auction. The results from the hypothetical and real WTP elicitations were then used to estimate calibration functions which were then used to adjust the hypothetical WTP values. The results of the study indicate that the differences between real and hypothetical WTP were not large, and only the difference in median WTP was significant. Still, no independent variables other than the individuals’ hypothetical WTP were significant in the calibration functions. The authors calculated ratios of the mean WTP estimated from the calibration models to the hypothetical mean WTP from the initial survey but as the ratios varied greatly their generalizability to another good is limited.

Cummings and Taylor (1999) suggested the use of the cheap talk approach which involves defining hypothetical bias for the respondents and explaining why it might happen so that respondents take this into account while answering to the valuation question. In all their experiments, there was no significant difference between the real and cheap talk WTP. Blackburn et al. (1994) suggested the use of a statistical bias function. In this, the authors identified individuals who misrepresented their true preferences then tested whether statistical bias functions including sociodemographic information such as gender, age and income could predict those individuals who would reverse their responses when going from the hypothetical to real WTP questions and those who would not. These responses would then be used to predict the real WTP for individuals from another sample or good. The study succeeded in sorting out the true yes-responses from the hypothetical yes-responses even though no individual independent variable was found to be significant at the 5% significance level. Overestimations and underestimations of WTP were also reported making it difficult to draw any definitive conclusions from the study. In a different experiment, Johannensson et al. (1999) estimated statistical bias functions based on the data from Johannensson et al. (1998) and Blumenschein et al. (1998). The statistical bias function contained a continuous variable regarding the individuals’ degree of certainty of their hypothetical WTP answer as well as a variable for the WTP price. The function correctly predicted 85% of the yes-yes and yes-no responses in the two experiments, and the results indicated that the higher the degree of certainty of the hypothetical yes-answer and the lower the price, the less risk of hypothetical bias.

The review authors concluded that substantial hypothetical bias is evident, and this can be context or good specific. They also noted that hypothetical bias does not depend on the elicitation method and also that it may be possible to calibrate the results so that the hypothetical WTP responses would mirror the actual behaviour of individuals. Of great significance, the review authors observe that the large differences in real and hypothetical WTP might be related to poor study design rather than to lack of validity for the CV method. Further, the authors highlight calibration methods concerning the DC method as: (i) Using a follow-up question on the respondent's degree of certainty (and only interpret the "absolutely sure" responses (or definitely sure depending on how the question is being phrased) as real yes-responses and; (ii) estimating a statistical bias function (takes the individuals' personal characteristics (where the degree of certainty could be one characteristic) and the WTP bid level into account when trying to sort out the real-yes responses). The authors note that these methods have not been used with any health or health care good, and that the goods used so far have been of low value. Further, these calibration methods have not been tested with a control group of real yes-responses, or indeed outside the experimental settings and therefore it is known whether the validity of the CV method increases with these methods or not. The authors conclude that the certainty of a hypothetical yes-response may be an important predictor of a real yes-response in WTP studies and therefore it may be possible to calibrate hypothetical WTP responses to better match real valuation.

*4. **List and Gallet (2001)***

List and Gallet (2001) reviewed the evidence on a range of mixed goods to provide evidence pertaining to the effects of various experimental protocol on the observed calibration factors. The authors made a pragmatic decision to focus only on studies that explicitly included discussion of experimental design variables that are commonly believed to affect stated preferences and therefore sought to answer the following six (6) questions in their review: (i) Does hypothetical bias exist in the typical contingent valuation exercise and if it does, what is the magnitude of the bias?; does hypothetical bias vary by (ii) WTP and WTA measures of value; (iii) elicitation methods; (iv) within-subject versus between subject experiments; (v) the use of field or laboratory experiments (vi) the distinction between public and private goods. The review is based on 29 studies which provided a total of 174 observations across both hypothetical and actual valuations. The studies included laboratory, field or both settings; public and private goods; within or between subject comparisons and different elicitation methods. For all review studies, a calibration factor was determined by dividing the mean hypothetical by mean actual values (H/A).

The authors find that the average person seems to exaggerate his or her actual WTP across a broad spectrum of goods with vastly different experimental protocol (Bohm 1972; Bishop & Heberlein 1979; Neill et al. 1994; Diamond et al. 1994; Fox et al. 1998; List & Shogren 1998; Balistreri et al. 2001). However, exceptions to this upward bias can be found(Johannesson et al. 1998) (Sinden, 1988). In their study, Sinden (1988) using a within group experiment in a laboratory setting elicited values for a public good using open ended questions. In the results, the ratios of hypothetical to actual WTP values across the experiments ranged from 0.80 – 1.50. Using a similar setting, Johannesson et al. (1988) compared hypothetical and actual values elicited for a private good using a dichotomous choice question. The authors used both within and between groups and had hypothetical to actual WTP ratios ranging from 0.88 to 1.33. Experimental results from the WTA literature were also mixed. In one study with two comparisons, WTA was both understated and overstated in comparisons of two related goods (goose licences and deer permits) (Bishop & Heberlein 1979). In one of the studies WTA understated real willingness to accept in the hypothetical regimes (List and Shogren (1999) while in the study by Smith and Mansfield's (1998) the two values are found to be statistically equivalent. Based on this narrative summary, the review authors conclude that hypothetical bias exists in contingent valuation exercises across a broad spectrum of goods.

As the relationship between the real and hypothetical stated values may be specific to experimental protocols, List and Gallet further investigated, through regression models, the impact of a range of variables: laboratory, field or both settings; public and private goods; whether WTP or WTA; within or between subject comparisons; and different elicitation methods. Review authors use three different regressand constructs: minimum, median, and maximum values of the calibration factor reported. Sample means within these three categories suggest that on average subjects overstate their preferences by a factor of about 3 in hypothetical exercises. However, the discrepancy between the minimum and maximum reported calibration factors is relatively small. The regression model shows hypothetical bias is not affected by (i) whether the experiment takes place in the lab or field or (ii) experimental designs, whether between-subject or within-subject. However, hypothetical bias is affected by: (i) elicitation methods (many of the theoretically incentive-compatible elicitation techniques do affect the calibration factor, suggesting that some methods induce more truthful responses than others); (ii) whether the respondent is providing WTA or WTP values (responses in WTP settings tend to correspond with actual WTP values more closely than hypothetical WTA values versus actual WTA stated values) and; (iii) the type of good - public or private good (hypothetical bias is considerably less for private goods compared to public goods). Based on this meta-analysis, the authors conclude that there is evidence that certain experimental protocol influence deviations in hypothetical and actual statements.

5. ***Little Joseph and Berrens Robert, 2004***

Little Joseph and Berrens Robert expand the original meta−analysis by List and Gallet (2001) using a significantly larger (29%) data set, and by including variables to account for referendum formats, certainty corrections, and cheap talk scripts. List and Gallet's (2001) criteria for the calculation of calibration factors (Hypothetical / Actual) and the general econometric approach were followed in examining the expanded dataset in this review. In addition to expanding the dataset, the authors explore weighting and clustering techniques since individual studies often produce multiple observations. Further, they conduct a probit model, where the dependent variable is the absence or presence of a statistically significant finding of bias between hypothetical and real stated values. Finally, both the extended calibration and probit models also include three new variables that aim to improve the credibility of hypothetical values. These variables account for the impact that referenda, cheap talk scripts, and certainty corrections may have on disparities between real and hypothetical valuation responses. In all cases use of the clustering correction (alone), provided the best fit.

Interpretation of the coefficients in this review is the same as that is the original meta-analysis by List and Gallet (2001). Results support the List and Gallet (2001) finding that the use of first price sealed bid auctions will reduce the disparity between hypothetical and real value. Contrary to the results in List and Gallet (2001), the results in this analysis indicate that the use of public goods referenda and certainty corrections will reduce the disparity between hypothetical and real values. Further, this analysis finds no evidence that the use of private goods will significantly reduce the disparity between real and hypothetical values. This indicates that calibration factors obtained from public goods referendum studies are lower than those obtained from non-referendum public goods studies. Further, estimates from the probit model show that use of certainty corrections will reduce the probability of observing a statistically significant disparity between real and hypothetical reported values (and the marginal effect of doing so can be large).

6. ***Murphy, J.J., Allen, P.G., Stevens, T.H. et al, 2005***

In this review, Murphy et al, 2005 have two objectives: (1) conduct a sensitivity analysis of the findings of List and Gallet (2001) meta-analysis of hypothetical bias in stated values and (2) conduct a meta-analysis using refined criteria. In their meta-analysis, List and Gallet (2001) conclude that the magnitude of hypothetical bias was statistically less for (a) WTP as compared to WTA applications, (b) private as compared to public goods, and (c) one elicitation method, the first price bid, as compared to the Vickery second-price auction baseline. In addressing the first objective, the authors in the current review begin by re-coding several observations in the List and Gallet (2001) meta-analysis. These include typing errors in observations that were reported incorrectly in their paper but were correct in the actual data used in their regressions, coding errors such as within group comparisons reported as between group comparisons in the Bohm (1972) study. Further, two observations that the current review authors could not be traced were dropped from the dataset.

After making these changes the current review authors re-estimated the model run by List and Gallet (2001) and found that although the changes affect the coefficient values, the results are qualitatively similar. However, Murphy et al. hypothesised that List and Gallet’s conclusions might have been driven not by the experimental protocol as reported but by results from a few influential studies. In the second stage, using the revised data which includes 29 studies (55 observations) they investigated the data for outliers and influential studies. Two of the WTA observations are from a single study(Cummings et al. 1986) with calibration factors that are at least 17 times greater than the mean of the other sis WTA studies. Further, this study used different mechanisms to elicit hypothetical and actual values (open-ended and Smith auction, respectively – it is possible that their calibration factors confound hypothetical bias with free-rider bias due to changing from a demand revealing mechanism to one that is not). After dropping these two observations, the model was re-run. In the results, private goods still produced a lower and statistically significant hypothetical bias than public goods but the WTP coefficient is no longer statistically significant.

A similar analysis conducted for the five elicitation mechanisms concluded that List and Gallet’s results were robust with respect to these changes. In a third revision, the authors adjusted the List and Gallet data for differences in interpretation. These included coding of studies which report values from different elicitation mechanisms in the hypothetical and actual settings and also a study which compared hypothetical

WTP to actual WTA. The authors purposed to keep only those studies which used the same elicitation mechanism for both the hypothetical and actual valuation, further reducing the data to 21 studies with a total of 32 observations. The results based on a re-run of the revised dataset showed that: (1) the statistically significant difference between WTP and WTA in the original LG results was sensitive to two extreme values that used different elicitation for actual and hypothetical valuation, and (2) private goods continued to have a lower bias than public goods. Further, the negative coefficients for lab experiments and within group comparisons were now weakly significant at the 10% level.

For the second objective, the authors reviewed 59 studies that reported both hypothetical and actual values. The dates covered in the review are not reported but the included papers span the years 1972 to 2003. To include an observation from a paper, (1) the hypothetical and actual values had to be elicited using the same mechanism; (2) Only WTP observations were included (authors argue that there are not enough WTA studies to truly capture important differences between WTP and WTA responses); and (3) hypothetical and actual vales had to be WTP measured in currency, not percentages of people responding “yes” to a dichotomous question. Dichotomous choice studies were only included if the authors provided an estimate of WTP. Following this criterion, the final dataset used in this review and meta-analysis included 28 studies yielding 83 observations.

Review authors assume that actual cash-based estimates are unbiased measures of the true WTP. For each observation, a calibration factor (CF) which is the ratio of hypothetical to actual value is calculated. The mean CF in this data is 2.60 but this comes from a highly skewed distribution with a 1.35 median CF. In this meta-analysis, the variables private and within group are defined the same was as defined in the List and Gallet meta-analysis. Owing to differences in the two reviews in the definition of lab, the current review authors create two new dummy variables, student and group, intended to capture essentially the same effects as List and Gallet’s lab variable. The student variable refers to whether the respondent is a student or otherwise while the group variable refers to the setting, not the nature of the decision. As there was high correlation between the student and group variables (pearson correlation coefficient = 0.77), the authors do not use both variables in the same model. As some elicitation mechanisms are typically associated with a particular type of good, the high correlation makes it difficult to isolate the effect of the elicitation mechanism from the type of good (Murphy et al. 2003). Based on this, the authors did not include dummy variables for each elicitation mechanism, choosing instead to create a new dummy variable that aggregates the elicitation mechanisms into two groups: choice = 1, for studies that use a choice-based elicitation method (e.g. dichotomous choice, polychotomous choice or payment card) and choice = 0 for the rest of the elicitation mechanisms. The authors also include a variable, calibrate =1 to indicate when an observation is based on any calibration technique, either instrument or statistical calibration.

In their first simple double log regression model that explains actual value as a function of the hypothetical value, the results indicate that for the range of hypothetical values in the sample, the bias increases as the hypothetical value increases. When evaluated at the mean hypothetical value (26.55), the predicted actual value is 10.24 which yield a calibration factor of 2.59; at the median hypothetical value (7.18), the predicted actual value is 3.89 with a 1.84 calibration factor.

The authors expanded the model to determine whether there are some factors that may help explain the cause of the bias. The expanded model adds the dummy variables for the type of respondent (student or not), good (private or public), type of comparison (within or between-group), type of elicitation mechanism (choice-based methods or otherwise) and whether the observation is based on any type of calibration technique (calibrate or not). With all the independent variables were evaluated at their means, the resulting predicted actual value was 8.83 and the CF is 3.01 while with the median, a CF of 2.47 is established. In this model, the coefficients for quadratic term for the natural log of the hypothetical value and within-group are both positive and significant. Contrary to List and Gallet's findings, the coefficient for private goods was not significant. Further, calibration techniques appear to be effective at reducing hypothetical bias while the positive and significant coefficient for choice indicates that the choice-based elicitation mechanisms are associated with less hypothetical bias. Finally, the negative coefficient on Student suggests that there may also be a subject pool effect. The authors suggest that since all the studies in the sample that use students are laboratory experiments, it is unclear whether the cause of hypothetical bias is the subject pool of the setting. In a second expanded model, the student variable was replaced with a “group” dummy variable (group=1 if values were elicited in a group setting such as a lab experiment, rather than an individual setting such as a phone or mail survey). With this model, the coefficient for group was negative and significant. Therefore, although there was clearly an effect, the authors state that they could not distinguish whether the cause of the difference was the subject pool or the setting. The calibrate model was not significant in this model while private which was not significant in the previous model was significant in this extended model, suggesting some sensitivity to model specification.

The authors also tested the sensitivity of their results to extreme values by dropping the five largest CFs. In the first of the estimations using the trimmed models, the independent variable included student and not group, with other independent variables remaining the same as in other models. The results of this model are consistent with those of the earlier model. The authors also test for the influence of studies with a large number of observations on hypothetical bias. They calculated the mean hypothetical and actual values from each study for a given set of independent variables. The resulting dataset had 45 observations with a mean CF of 3.26 (median: 1.50). Based on the results of this new model, the hypothetical value seems to be the best predictor of actual value.

Based on the results of this review, Murphy et al (2005) note that: (i) the hypothetical value seems to be the best predictor of actual value and (ii) calibration techniques are effective at reducing hypothetical bias. Further, the authors observe that as with previous meta-analysis, a meta-analysis of hypothetical bias appears to be very sensitive to: (a) model specification; (b) a lack of variability in the data and; (c) treatment of extreme values. The authors conclude that hypothetical bias in SP studies may not be as important as most previous studies suggest. Further, they question the prevailing wisdom about several of the factors responsible for this bias.

Appendix 10: Sample search strategy for criterion validity systematic review

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	2,393
S9	TI Will* AND Accept or WTA	250
S8	TI Will* AND Pay or WTP	2,274
S7	S3 AND S6	2
S6	S4 AND S5	1,229
S5	TI Actual OR revealed OR real OR inconsequentiality OR Direct	46,914
S4	TI Stated OR Hypothetical OR Contingent OR Consequentiality OR indirect	7,109
S3	S1 AND S2	6,622
S2	TI Validity OR Valid*	33,026
S1	TI External OR Criterion OR Predictive OR Reliability	34,557

Appendix 11: Criterion validity systematic review data extraction terms

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
Money given (for participation in surveys or purchase of valuation good)	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	Author conclusions on validity
Duration between surveys	Reasons given for disparity in hypothetical and actual values
Validity assessment method	

Appendix 12: Criterion validity assessment hypotheses and definition of study attributes

Criterion validity assessment hypotheses

A rational consumer is expected to maximize the utility he or she obtains from the consumption of a commodity subject to their budget constraints and this directly determines their willingness to pay (WTP) for the commodity. WTP is a function of income and a vector of prices faced by the individual and the alternative levels of the good or quality indexes.

Utility depends on a vector of individual characteristics influencing the trade-off that the individual is prepared to make between income and the attributed of the good. Consumers have a utility function u(*) where u is the consumer's utility and * represents all other factors expected to influence the utility derived from the consumption of a given commodity and may include the quality of the commodity and personal attributed such as perception of the commodity and need, among others. This consumer also minimizes expenditures subject to a utility constraint, u=u*. His expenditure function, e(p,u*), results where e(*) is the minimum amount of expenditures necessary to produce u*, and p is the price of x.

The criterion validity of WTP values has been questioned, with critics arguing that hypothetical WTP values overestimate actual values. The assessment of the criterion validity of WTP involves the comparison of values obtained in a hypothetical setting with those derived from an actual survey. While there is no known economic theory to guide this assessment, some economic theory and evidence from prior criterion validity tests provide a framework within which further assessments can be conducted. The current meta-regression is based on the following two broad hypotheses on variables that are expected to influence higher or lower WTP ratios. The justification for the use of the different variables is also provided.

Hypotheses Group 1: The ratio of hypothetical to actual WTP is expected to be higher when:

1) A public good is valued.
2) Student or non-users of the valuation good are used.
3) Different samples are used in hypothetical and actual surveys.

4) Payment for participation in the survey.
5) Duration of more than 2 weeks between the hypothetical and actual surveys.
6) Elicitation format (Use of non-DC methods).
7) Non– personal study administration modes are used.

Hypotheses Group 2: The ratio of hypothetical to actual WTP is expected to be lower when:

1) Private goods are valued
2) Non-student samples and potential users of the valuation good are used
3) Same samples are used in both surveys
4) The duration between the hypothetical and actual surveys is less than 2 weeks
5) Elicitation format (Dichotomous choice questions are used)
6) In-person interviews are used

Definition and justification for variables that will be tested in the hypotheses.

<u>Class of good (Public versus private and quasi-public)</u>

The nature of public goods is that regardless of their financing and provision, they are non-rival in consumption and non-excludable. As the majority of public goods are already provided by the government, individuals may not act rationally in their valuation in both hypothetical and actual surveys. Public goods are also characterised by free riding. Individuals may give very high estimates in the hypothetical setting to influence their provision but faced with the actual payment decision they may provide low values with the knowledge that these would be provided regardless of the amount they pay for them. When using DC questions to value these, in the hypothetical survey a large number of individuals may be willing to pay for them for the same reason but only a fraction of these pay in the actual survey because they understand that they would access the goods regardless of their payment decision. Valuation of public goods is also an unfamiliar process for most individuals, for the reasons indicated earlier. Previous studies have found that the variation between hypothetical and actual surveys is higher when public goods are valued, compared to private goods (List and Gallet, 2001).

Student samples

In most cases, student samples are composed of individuals who are unlikely to be the market for the valuation commodity. Further, they may not have the financial outlay to transact at the time of the survey and may therefore make their decisions with the knowledge that their decisions are not consequential. When the surveys are held in non-field settings such as classrooms the reality of the exercise is absent. Combined, these factors would lead to valuation decisions which are not rationally determined and hence huge variations in hypothetical to actual WTP values. The use of student samples has also been identified as a source of hypothetical bias in previous criterion validity assessments(Murphy et al. 2005).

Within or between samples

Studies have found that within sample designs are appropriate for the assessment of criterion validity as they offer the potential for the researchers to control for individual-specific effects in the statistical analysis (List & Gallet 2001).

Payment for participation in the survey

Payment for participation in a survey or for the purchase of a valuation commodity during a survey is expected to influence an individual's decision. In the first instance, provision of money takes away the consequentiality from the decision-making process. The individual may thus make irrational valuation decisions. One of these may be stating high hypothetical values but lower actual values. Secondly, the individual may regard this as the opportunity cost of participating in the survey and still not reveal their true preferences in one or both of the surveys, leading to variations in the values obtained. Failure to provide money for participation in the surveys may mean opportunity costs for the individual, interfering with their budget which then influences the valuation responses that they provide.

Duration between hypothetical and actual surveys

Psychologists argue that a 2 weeks duration between surveys (test – retest) is expected to provide closest estimates(Carmines & Zeller 1979). Further, majority of the factors that influence the individual's demand for the commodity such as income, preferences, prices and even personal factors such as health status are not expected to have changed. A period of more than 2 weeks may lead to poor recall of stated values or choices and other variables might not hold constant for longer than this time too resulting in variation between hypothetical and actual values.

Elicitation formats

The DC elicitation format resembles the normal pricing of goods that individuals are familiar with in most markets and so they may easily recall of identify the thought process that they undertake to arrive at everyday purchase decisions. When other unfamiliar methods such as open-ended methods are used then individual may not accurately estimate their WTP in either survey. Also, depending on the duration between the two surveys it may be easier for the individual to recall a Yes or No response to a DC question as opposed to the response provided for an open-ended question. Previous meta-analysis have found that the elicitation format influences the ratio of hypothetical to actual WTP values significantly (Little & Berrens 2004).

Study administration modes

Non-personal modes in this analysis include mail and internet surveys. In person interviews, including telephone surveys provide an opportunity for clarification of the hypothetical scenarios and the valuation exercise. While this may introduce bias, they also provide opportunities for the interviewer to engage individuals in a discussion on the exercise and remind them of their budget constraints, making the situation as real as possible. Mail surveys have been shown in the current review to have the lowest response rates and are associated with high hypothetical bias.

Appendix 13: Background characteristics of papers included in the review

No.	Reference	#. of comparisons	Country	Sector	Class of good	Validity term	Ratio/Odd ratio (Hypothetical/ Actual WTP)	Author conclusion on criterion validity (# comparisons)
1	(Balistreri et al. 2001)	1	USA	Other	Pure Private	Non-Specific	1.25	Confirmed
2	(Bhatia & Fox-Rushby 2003)	1	India	Health	Pure Private	Criterion	0.94	Confirmed
3	(Bishop & Heberlein 1979)	1	USA	Environment	Quasi-Private	Non-Specific	1.60	Not Confirmed
4	(Blumenschein et al. 2008)	6	USA	Health	Pure Private	Non-Specific	0.89 – 2.00	Not Confirmed (6)
5	(Blumenschein et al. 1998)	2	USA	Other	Pure Private	Hypothetical bias	9.1 13.63	Confirmed (1) Not Confirmed (1)
6	(Blumenschein et al. 2001)	5	USA	Health	Pure Private	Non-Specific	2.07 – 3.68	Not confirmed (4)
7	(Botelho & Pinto 2002)	1	Portugal	Environment	Quasi-Private	Hypothetical bias	11.50	Not Confirmed
8	(Bratt 2010)	3	El Salvador, Egypt	Health	Pure Private	Criterion	0.94 – 1.21	Not Confirmed (3)
9	(Brown et al. 1996)	2	USA	Environment	Pure Public	Non-Specific	4.10 – 6.44	Not Confirmed (2)
10	(Brown & Taylor 2000)	2	USA	Environment	Pure Public	Non-Specific	8.65 – 11.76	Not Confirmed (2)
11	(Bryan & Jowett 2010)	1	UK	Health	Pure Private	Hypothetical bias	-	Confirmed
12	(Byrnes et al. 1999)	2	USA	Environment	Quasi-Private	Non-Specific	9.05 – 10.1 8	Not Confirmed (2)
13	(Camacho-Cuena et al. 2004)	4	Spain	Other	Quasi-Private	Hypothetical bias	0.93 – 1.06	Confirmed (4)
14	(Carlson 2000)	3	USA	Other	Pure Private	Criterion	1.30 – 3.34	Not Confirmed (3)
15	(Champ & Bishop 2001)	1	USA	Environment	Quasi-Private	Non-Specific	1.71	Not Confirmed
16	(Champ et al. 1997)	2	USA	Environment	Pure Public	Non-Specific	4.10 – 6.44	Not Confirmed (2)
17	(Cummings et al. 1995)	5	USA	Other	Pure Private	Non-Specific	2.56 – 10.5	Not Confirmed (5)
18	(Cummings et al. 1997)	1	USA	Environment	Quasi-	Non-Specific	1.67	Not Confirmed

No.	Reference	#. of comparisons	Country	Sector	Class of good	Validity term	Ratio/Odd ratio (Hypothetical/ Actual WTP)	Author conclusion on criterion validity (# comparisons)
					Private			
19	(Fox et al. 1998)	2	USA	Health	Pure Private	Non-Specific	1.48 – 1.69	Not Confirmed (2)
20	(Frykblom 1997)	3	Sweden	Other	Pure Private	Non-Specific	1.49 – 1.77	Not Confirmed (3)
21	(Getzner 2000)	1	Austria	Environment	Pure Public	Hypothetical bias		Not Confirmed
22	(Heberlein & Bishop 1986)	6	USA	Other	Quasi-Private	Non-Specific	1.24 – 2.98 0.70 – 1.60	Confirmed (3) Not Confirmed (3)
23	(Johannesson 1997)	1	Sweden	Other	Pure Private	Hypothetical bias	1.633	Not Confirmed
24	(Johannesson et al. 1997)	1	Sweden	Other	Pure Private	Non-Specific	1.02	Confirmed
25	(Johannesson et al. 1998)	4	Sweden	Other	Pure Private	Non-Specific	0.80 – 10.29	Not Confirmed (4)
26	(Johnston 2006)	1	USA	Environment	Quasi-Private	Criterion	1.08	Confirmed
27	(List 2001)	4	USA	Other	Pure Private	Hypothetical bias	1.02 – 1.94	Not Confirmed (4)
28	(List & Shogren 2002)	1	USA	Other	Pure Private	Non-Specific	0.69	Not Confirmed
29	(List & Shogren 1998)	3	USA	Other	Pure Private	Non-Specific	2.18 – 3.47	Not Confirmed (3)
30	(Loomis et al. 1996b)	2	USA	Other	Pure Private	Criterion	1.95 – 3.64	Not Confirmed (2)
31	(Loomis et al. 1997)	4	USA	Other	Pure Private	Non-Specific	1.85 – 3.00	Not Confirmed (2)
32	(Macmillan et al. 1999)	1	USA	Environment	Pure Public	Non-Specific	0.91	Confirmed
33	(Mozumder & Berrens 2007)	2	USA	Other	Pure Public	Hypothetical bias	0.99 – 1.00	Not Confirmed (2)
34	(Muller & Ruffieux 2011)	1	France	Other	Pure Private	External	0.98	Confirmed
35	(Murphy et al. 2002)	4	USA	Environment	Pure Public	Hypothetical bias	2.43 – 7.57	Not Confirmed (4)
36	(Murphy et al. 2010)	9	USA	Other	Pure Private	Hypothetical bias	0.94 – 2.12	Not Confirmed (9)
37	(Neill et al. 1994)	2	USA	Other	Pure Private	Non-Specific	3.90 – 25.08	Not Confirmed (2)
38	(Onwujekwe et al. 2001)	6	Nigeria	Health	Pure Private	Criterion	0.92 – 1.42	Confirmed (6)
39	(Onwujekwe &	2	Nigeria	Health	Pure Private	Criterion		Not confirmed (2)

No.	Reference	#. of comparisons	Country	Sector	Class of good	Validity term	Ratio/Odd ratio (Hypothetical/ Actual WTP)	Author conclusion on criterion validity (# comparisons)
	Uzochukwu 2004)							
40	(Onwujekwe 2004a)	3	Nigeria	Health	Pure Private	Hypothetical bias		Not Confirmed (3)
41	(Onwujekwe et al. 2005)	3	Nigeria	Health	Pure Private	Criterion		Not Confirmed (3)
42	(Onwujekwe 2001a)	2	Nigeria	Health	Pure Private	Predictive	1.23 – 1.40	Confirmed (2)
43	(Paradiso & Trisorio 2001)	2	UK	Other	Pure Private	Non-Specific	2.79 – 3.45	Not Confirmed (2)
44	(Ramke et al. 2009)	3	East Timor	Other	Pure Private	Criterion	1.06 1.96 – 2.63	Confirmed (1) Not Confirmed (2)
45	(Seip Strand, J. 1992)	1	Norway	Environment	Pure Public	Non-Specific	6.72	Not Confirmed
46	(Veisten & Navrud 2006)	34	Norway	Environment	Pure Public	Non-Specific	1.01 – 18.03	Not Confirmed (34)
47	(Vernazza et al. 2015b)	2	UK, Germany	Health	Pure Private	Non-Specific		Not Confirmed (2)
48	(Vossler, Ethier, et al. 2003)	2	USA	Environment	Quasi-Private	Criterion	0.003 – 0.004	Not Confirmed (2)
49	(Vossler & Kerkvliet 2003)	3	USA	Environment	Pure Public	Criterion	1.01 -	Confirmed (3)
50	(Willis & Powe 1998)	1	UK	Environment	Quasi-Private	Criterion	110.82	Not Confirmed

Appendix 14: Background characteristics of papers included in the meta-analysis and meta-regression

No	Reference	No. of comparisons	Country	Sector	Class of good	Summary measure	Meta-analysis / Meta-regression*	Ratio/Odd ratio range (Hypothetical/ Actual WTP)	Author conclusion on criterion validity *
1	(Balistreri et al. 2001)	1	USA	Other	Pure Private	Mean	Both	1.25	Confirmed
2	(Bhatia & Fox-Rushby 2003)	1	India	Health	Pure Private	Percentage	Both	0.94	Confirmed
3	(Bishop & Heberlein 1979)	1	USA	Environment	Quasi-Private	Mean	Meta-regression only	1.60	Not Confirmed
4	(Blumenschein et al. 2008)	6	USA	Health	Pure Private	Mean	Both	0.89 – 2.00	Not Confirmed (6)
5	(Blumenschein et al. 1998)	2	USA	Other	Pure Private	Percentage	Both	9.1 13.63	Confirmed (1) Not Confirmed (1)
6	(Blumenschein et al. 2001)	4	USA	Health	Pure Private	Mean (1) Percentage (3)	Meta-regression only (1) Both (3)	2.07 – 3.68	Not confirmed (4)
7	(Botelho & Pinto 2002)	1	Portugal	Environment	Quasi-Private	Mean	Both	11.50	Not Confirmed
8	(Bratt 2010)	3	El Salvador, Egypt	Health	Pure Private	Percentage	Both	0.94 – 1.21	Not Confirmed (3)
9	(Brown et al. 1996)	2	USA	Environment	Pure Public	Mean	Both	4.10 – 6.44	Not Confirmed (2)
10	(Brown & Taylor 2000)	2	USA	Environment	Pure Public	Mean	Both	8.65 – 11.76	Not Confirmed (2)
12	(Byrnes et al. 1999)	2	USA	Environment	Quasi-Private	Mean	Meta-regression	9.05 – 10.1 8	Not Confirmed (2)

No	Reference	No. of comparisons	Country	Sector	Class of good	Summary measure	Meta-analysis / Meta-regression*	Ratio/Odd ratio range (Hypothetical/ Actual WTP)	Author conclusion on criterion validity *
							only		
13	(Camacho-Cuena et al. 2004)	4	Spain	Other	Quasi-Private	Mean	Both	0.93 – 1.06	Confirmed (4)
14	(Carlson 2000)	3	USA	Other	Pure Private	Mean	Both	1.30 – 3.34	Not Confirmed (3)
15	(Champ & Bishop 2001)	1	USA	Environment	Quasi-Private	Mean	Both	1.71	Not Confirmed
16	(Champ et al. 1997)	2	USA	Environment	Pure Public	Mean	Both	4.10 – 6.44	Not Confirmed (2)
17	(Cummings et al. 1995)	5	USA	Other	Pure Private	Percentage	Both	2.56 – 10.5	Not Confirmed (5)
18	(Cummings et al. 1997)	1	USA	Environment	Quasi-Private	Percentage	Both	1.67	Not Confirmed
19	(Fox et al. 1998)	2	USA	Health	Pure Private	Mean	Meta-regression only	1.48 – 1.69	Not Confirmed (2)
20	(Frykblom 1997)	3	Sweden	Other	Pure Private	Mean Percentage	Both	1.49 – 1.77	Not Confirmed (3)
23	(Johannesson 1997)	1	Sweden	Other	Pure Private	Mean	Both	1.633	Not Confirmed
24	(Johannesson et al. 1997)	1	Sweden	Other	Pure Private	Mean	Both	1.02	Confirmed
25	(Johannesson et al. 1998)	4	Sweden	Other	Pure Private	Mean	Both	0.80 – 10.29	Not Confirmed (4)
26	(Johnston 2006)	1	USA	Environment	Quasi-Private	Mean	Both	1.08	Confirmed

No	Reference	No. of comparisons	Country	Sector	Class of good	Summary measure	Meta-analysis / Meta-regression*	Ratio/Odd ratio range (Hypothetical/ Actual WTP)	Author conclusion on criterion validity *
27	(List 2001)	4	USA	Other	Pure Private	Mean	Both	1.02 – 1.94	Not Confirmed (4)
28	(List & Shogren 2002)	1	USA	Other	Pure Private	Mean	Both	0.69	Not Confirmed
29	(List & Shogren 1998)	3	USA	Other	Pure Private	Mean	Both	2.18 – 3.47	Not Confirmed (3)
30	(Loomis et al. 1996b)	2	USA	Other	Pure Private	Mean	Both	1.95 – 3.64	Not Confirmed (2)
31	(Loomis et al. 1997)	4	USA	Other	Pure Private	Mean	Both	1.85 – 3.00	Not Confirmed (2)
32	(Macmillan et al. 1999)	1	USA	Environment	Pure Public	Mean	Both	0.91	Confirmed
33	(Mozumder & Berrens 2007)	2	USA	Other	Pure Public	Mean	Meta-regression only	0.99 – 1.00	Not Confirmed (2)
34	(Muller & Ruffieux 2011)	1	France	Other	Pure Private	Mean	Meta-regression only	0.98	Confirmed
35	(Murphy et al. 2002)	4	USA	Environment	Pure Public	Mean	Meta-regression only	2.43 – 7.57	Not Confirmed (4)
36	(Murphy et al. 2010)	9	USA	Other	Pure Private	Mean	Both	0.94 – 2.12	Not Confirmed (9)
37	(Neill et al. 1994)	2	USA	Other	Pure Private	Mean	Both	3.90 – 25.08	Not Confirmed (2)
38	(Onwujekwe et al. 2001)	6	Nigeria	Health	Pure Private	Percentage	Both	0.92 – 1.42	Confirmed (6)
42	(Onwujekwe 2001a)	2	Nigeria	Health	Pure Private	Percentage	Both	1.23 – 1.40	Confirmed (2)

No	Reference	No. of comparisons	Country	Sector	Class of good	Summary measure	Meta-analysis / Meta-regression*	Ratio/Odd ratio range (Hypothetical/ Actual WTP)	Author conclusion on criterion validity *
43	(Paradiso & Trisorio 2001)	2	UK	Other	Pure Private	Mean	Both	2.79 – 3.45	Not Confirmed (2)
44	(Ramke et al. 2009)	3	East Timor	Other	Pure Private	Percentage	Both	1.06 1.96 – 2.63	Confirmed (1) Not Confirmed (2)
45	(Seip Strand, J. 1992)	1	Norway	Environment	Pure Public	Percentage	Both	6.72	Not Confirmed
46	(Veisten & Navrud 2006)	34	Norway	Environment	Pure Public	Mean (6), Percentage (27),	Both (33)	1.01 – 18.03 -	Not Confirmed (34)
48	(Vossler, Ethier, et al. 2003)	2	USA	Environment	Quasi-Private	Percentage	Both	0.003 – 0.004	Not Confirmed (2)
49	(Vossler & Kerkvliet 2003)	3	USA	Environment	Pure Public	Percentage (1)	Both (1)	1.01 -	Confirmed (3)
50	(Willis & Powe 1998)	1	UK	Environment	Quasi-Private	Percentage	Both	110.82	Not Confirmed

***Number of comparisons**

Appendix 15: Definition of variables included in meta-regression models

Variable name	**Description**
Log ratio	Natural log of the ratio of hypothetical to actual WTP values (for mean summaries)
Log Odds ratio	Natural log of the odds ratio of hypothetical to actual WP values (for percentage summaries)
Country income level	Study conducted in high income county = 1; Otherwise = 0
Sector	Sector as categorical variable: Reference category (Health); Environment, Other *Sector categories as binary variables* Study in Environment sector = 1; Otherwise = 0 Study in Health sector = 1; Otherwise = 0 Study in Other sector = 1; Otherwise = 0
Class of good	Class of good as categorical variable: Reference category (Pure Private); Pure Public, Quasi-Private *Class of good categories as binary variables* Pure Public good = 1; Otherwise = 0 Pure Private good = 1; Otherwise = 0 Quasi-Private good = 1; Otherwise = 0
Purpose of good	Purpose of good as categorical variable: Reference category (Prevention); Conservation, Other *Purpose of good as binary variables* Conservation purpose = 1; Otherwise = 0 Prevention purpose = 1; Otherwise = 0 Other purpose = ; Otherwise = 0
Duration between surveys	Hypothetical and actual surveys held concurrently = 1; Otherwise = 0
Payment duration	One-off payment elicited = 1; Otherwise = 0

Variable name	Description
Payment Vehicle	Cash fee elicited = 1; Otherwise = 0
Type of comparison	Between sample comparisons = 1; Otherwise = 0
Survey setting	Surveys held in field setting = 1; Otherwise = 0
Money effects (1)	Money given for participation in survey = 1; Otherwise = 0
Money effects (2)	Money given in actual survey for purchase of valuation good = 1; Otherwise = 0
Comparisons between hypothetical and actual surveys	
Student sample	Respondents in both surveys were students = 1; Otherwise = 0
Users	Users or potential users in both surveys = 1; Otherwise = 0
Sample selection	Same selection method in both surveys = 1; Otherwise = 0
Sample selection categories	Random sampling in both surveys = 1; Otherwise = 0 Purposive sampling in both surveys = 1; Otherwise = 0 Convenience sampling in both surveys = 1; Otherwise = 0
Administration mode	Same administration mode used in both surveys = 1; Otherwise = 0
Administration mode categories	Mail administration in both surveys = 1; Otherwise = 0 In-person administration in both surveys = 1; Otherwise = 0
Elicitation method	Same elicitation method used in both surveys = 1; Otherwise = 0
Elicitation mode categories	Auction method in both surveys = 1; Otherwise = 0 Bidding method in both surveys = 1; Otherwise = 0 Dichotomous choice method in both surveys = 1; Otherwise =0 Open ended methods in both surveys = 1; Otherwise = 0

Appendix 16: Descriptive statistics for variables included in the meta-regression models

Variable name	**Value labels**	**Mean Summaries**	**Percent Summaries**
		# of comparisons* (# of studies)	**# of comparisons* (# of studies)**
Country income level	High Income country	84 (32)	41 (8)
	Other income categories	0	15 (5)
Sector	Health	9 (6)	15 (5)
	Environment	23 (11)	30 (5)
	Others	52 (18)	11 (4)
Class of good	Pure public good	22 (8)	26 (3)
	Pure private good	42 (18)	26 (8)
	Quasi-private good	20 (8)	4 (3)
Purpose of good	Conservation	23 (10)	28 (3)
	Prevention	9 (3)	19 (7)
	Other	52 (19)	9 (4)
Duration between surveys	Concurrent	76 (29)	25 (6)
	Non-concurrent	8 (4)	31 (8)
Payment duration	One-off duration	82 (30)	52 (11)
	Other durations	2 (2)	4 (2)
Payment Vehicle	Cash-fee payments	67 (27)	29 (11)
	Other payment vehicles	17 (6)	27 (2)
Type of comparison	Between-Study	53 (22)	11 (5)
	Within-Study	31 (11)	45 (10)
Survey setting	Field	29 (11)	50 (12)
	Laboratory	55 (21)	6 (3)
Money effects (1)	Money given in both surveys	70 (29)	34 (5)

Variable name	Value labels	Mean Summaries	Percent Summaries
		# of comparisons* (# of studies)	# of comparisons* (# of studies)
Money effects (2)	Respondents given cash in actual survey for purchase	14 (3)	1 (1)
Comparisons between hypothetical and actual surveys			
Student sample	Students in both surveys	36 (13)	50 (12)
	Non-students	48 (19)	6 (3)
Users	Respondents users / potential users of valuation good	75 (28)	54 (12)
	Non-users	9 (4)	2 (2)
Sample selection	Same method in both surveys	13 (6)	52 (11)
	Different methods	71 (26)	4 (3)
Sample selection categories	Random sampling in both surveys	4 (3)	15 (6)
	Convenience sampling in both surveys	25 (7)	31 (4)
	Purposive sampling in both	42 (16)	6 (3)
Administration mode	Same mode on both surveys	77 (28)	52 (11)
	Different modes	7 (4)	4 (2)
Administration mode categories	Mail administration in both surveys	19 (7)	24 (1)
	In-person surveys in both surveys	58 (21)	28 (10)
Elicitation method	Same method in both surveys	58 (22)	52 (10)
	Different methods	26 (13)	4 (3)
Elicitation mode categories	Auction methods	23 (11)	-
	Bidding methods	1 (1)	5 (1)
	Dichotomous methods	25 (12)	38 (10)
	Open ended methods	14 (8)	9 (1)

* The number of comparisons higher than the number of studies because some studies generated multiple estimates

Appendix 17: Meta-regression results for mean summaries with dummy variables for sector, purpose and class of good

	Base model		**Reduced model**
Variables	Coefficient (SE)		Coefficient (SE)
General study attributes			
Sector (Reference: Health Sector)		*Sector (Reference: Environment Sector)*	
Environment Sector	0.000 (0.535)	Health Sector	-1.291*** (0.258)
Other Sector	-0.450 (0.322)	Other Sector	-1.371*** (0.191)
Class of good (Reference: Pure private good)			
Pure Public Good	-0.198 (0.299)		-
Quasi-Private Good	0.0260 (0.254)		-
Purpose of good (Reference: Prevention)			
Conservation Purpose	1.030** (0.402)		-
Other Purposes (Besides, Prevention and Conservation)	0		-
Duration between surveys			
Hypothetical and Actual surveys held concurrently	0.291(0.230)		-
Payment duration			
One-off payment elicited	1.178*** (0.390)		1.653*** (0.517)
Payment vehicle			
Cash fee payment vehicle	-0.234 (0.300)		-
Type of comparison			
Between sample comparisons	-0.160 (0.202)		-
Survey setting			
Surveys held in a field setting	-0.267 (0.343)		-
Money effects			
Money given for participation in either survey	-0.090 (0.190)		-
Comparisons between study attributes in hypothetical and actual surveys			
Sample type [Student or not]			
Student sample in both surveys	0.425 (0.377)		-
Sample type [Potential user or not]			
Respondent a potential user in both surveys	0.482* (0.268)		0.278* (0.144)
Same sample selection method in both surveys			
Random sampling in both	-1.581*** (0.475)		-1.220*** (0.315)
Purposive sampling in both	-0.255 (0.276)		-

	Base model		**Reduced model**
Variables	Coefficient (SE)		Coefficient (SE)
Convenience sampling in both	-0.407 (0.349)		-
Administration mode			
Mail administration in both surveys	-0.873** (0.355)		-0.826*** (0.229)
In-person surveys in both surveys	-0.780*** (0.284)		-0.528** (0.205)
Elicitation method			
Auction method in both surveys	0.877** (0.431)		0.720*** (0.182)
Bidding method in both surveys	0.459 (0.745)		-0.768*** (0.186)
Dichotomous choice methods in both surveys	0.425 (0.272)		0.311** (0.155)
Open ended methods in both surveys	0.660** (0.297)		0.587*** (0.138)
Payment card method in both surveys	-0.070 (0.254)		-
Constant	-0.254 (2.267)		-0.0693 (1.265)
Observations	84		84
R-squared	0.672		0.633

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

Appendix 18: Meta-regression results for percent summaries with dummy variables for sector, purpose and class of good

	Base model	**Reduced model**
Variables	Coefficient (SE)	Coefficient (SE)
General study attributes		
High Income Country	1.919**(0.749)	1.546***(0.482)
Sector (Reference: Health Sector)		
Environment Sector	-1.143(0.818)	-
Other Sector	0.401(0.268)	-
Class of good (Reference: Pure public good)		
Pure Private Good	10.96***(1.088)	-
Quasi-Private Good	-	-
Purpose of good (Reference: Conservation)		
Prevention Purpose	-9.996***(0.821)	-
Other Purposes (Besides, Prevention and Conservation)	-0.731(0.669)	-
Duration between surveys		
Hypothetical and Actual surveys held concurrently	-0.503(0.381)	-
Payment duration		
One-off payment elicited	-	-
Payment vehicle		
Cash fee payment vehicle	-3.856***(1.406)	
Type of comparison		
Between sample comparisons	-0.393(0.614)	-2.069***(0.640)
Survey setting		
Surveys held in a field setting	-	
Money effects		
Money given for participation in either survey	0.0696 (0.485)	-
Comparisons between study attributes in hypothetical and actual surveys		
Sample type [Student or not]		
Student sample in both surveys	5.199***(1.405)	-
Sample type [Potential user or not]		
Respondent a potential user in both surveys	0.726(0.617)	
Same sample selection method in both surveys		
Random sampling in both	4.630***(1.069)	-1.498**(0.625)

	Base model	**Reduced model**
Variables	Coefficient (SE)	Coefficient (SE)
Purposive sampling in both	4.500***(1.060)	
Convenience sampling in both	-	-
Administration mode		
Mail administration in both surveys	-4.287***(1.445)	
In-person surveys in both surveys	-	1.775***(0.580)
Elicitation method		
Bidding method in both surveys	-5.198***(1.089)	
Dichotomous choice methods in both surveys	-5.127***(1.087)	
Open ended methods in both surveys	-4.724***(1.558)	
Constant	10.84**(5.139)	1.666**(0.730)
Observations	56	56
R-squared	0.871	0.488

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

Appendix 19: The Malaria WTP study context

Introduction

The republic of India, in South Asia covers an area of 3.287km^2, making it the seventh-largest country by area. India is the second-most populous country and democracy in the world with an estimated population of 1.311 billion (2015)[45](UN 2017). Geographically, the country is composed of six physiographic regions ranging from the northern mountainous regions which include the Himalayas of Nepal to the Peninsular Plateau, the Great Plains, Thar Desert and the eastern coastal plain. The country is divided into twenty-nine (29) states and 7 union territories. The states are further subdivided into districts, talukas (sub-districts) and villages for ease of administration. Each village elects its representatives and decided on their priorities and development. The district is considered to be a very important administrative unit in India and is responsible for law and order, revenue, justice, facilities and service to the people.

Climate

There are two main seasons in the country: the hot and wet season which runs from mid-June to mid-September with the peak in May and the dry season which runs from mid-December to February with the peak in January. Further, there are two transitional seasons in the country – the pre-monsoon hot and dry season from March to mid-June and the season of the retreating monsoon from mid-September to December.

Malaria in India

Malaria has been a major public health problem in the country for centuries. Early estimates of malaria prevalence[46] in India in 1947 indicated that one in every five people suffered from an episode of malaria with an annual incidence[47] of 75 million and 0.8 million deaths (Lal et al. 2000; Kumar et al. 2007). The country experiences variable malaria endemicity[48] from hypo endemic[49] to hyper endemic[50] conditions with the

[45]Total area excludes disputed territories not under Indian control

[46] Prevalence *refers to* the proportion of people infected with malaria at a given point in time (WHO 2008)

[47]Incidence refers to the number of newly diagnosed malaria cases during a defined period in a specified population

[48] Endemicity (or disease intensity) is a measure of the degree of malaria transmission in an area (Beljaev et al. 2016)

[49] Hypoendemic: Very intermittent malaria transmission/less than 10% transmission in the 2-9 year group (Beljaev et al. 2016)

Northwest regions periodically facing widespread fulminant epidemics. The north western regions of the Brahamputra valley and a narrow strip on the western coast of South India reportedly experienced no malaria incidences (*India 1958* in Misra 1999).

Malaria parasites and transmission

There are four human malaria parasite species and only three of these exist in India. Two of these, *P. vivax* and *P. falciparum*, are prevalent all over the country. The *P. falciparum* species predominates in hilly and foothill areas of the country where the transmission season is longer, or more than one vector transmits malaria. The third species, *P. malariae*, constitutes less than 1% of the parasites and is found in highly stable ecosystems inhabited by the tribal people of India. The malaria transmission season differs across the different regions in the country depending on the temperature and rainfall patterns. In the northern region transmission occurs between May and October, starting one month earlier than the north eastern states while in the southern states transmission is in the later part of the year (*Sharma 1986* in Misra 1999).

The government of India estimates that about 95% population in the country resides in malaria endemic areas. Further, an estimated 80% of malaria reported in the country is confined to areas consisting 20% of population residing in tribal, hilly, difficult and inaccessible areas.

Malaria control initiatives

To combat the growing incidence of malaria cases, the government of India launched the National Malaria Control Programme (NCMP) in 1953. This was largely as a result of successful pilot studies in various parts of the country where indoor residual spraying (IRS) with dichloro-diphenyl-trichlorethane (DDT) chemical was conducted. The chemical was primarily sprayed in human dwellings and cattle sheds with limited intervention in hospitals and clinics. The NCMP proved highly successful with the recorded cases of malaria dropping to 2 million by 1958 from estimated 75 million in 1947 (NVBDCP) (India 1976). Within this 5-year period, child spleen, child parasite and infant parasite rates were reduced by 73.2%, 80% and 62.4% respectively (Misra 1999). With this success, the programme was changed in 1958 to a more ambitious and time bound programme, the National Malaria Eradication Programme (NMEP) whose vision

[50] Hyperendemic: Intense transmission, but with periods of no transmission during dry season / 51-75% transmission in the 2-9 year group (Beljaev et al. 2016)

was to achieve total coverage with IRS along with case detection and treatment, with the aim of eradicating malaria. With the new programme, by 1961 the incidence dropped further to a mere 49151 cases, with no deaths (Dash et al. 2008). The NMEP faced repeated technical, operational and administrative challenges, compounded by DDT shortages in the 1960s and 1970s which led to a resurgence of malaria in the mid-seventies with 6.45 million cases reported in 1975 (Pattanayak et al. 1994). The implementation of the urban malaria scheme (UMS) in 1971-72 and the modified plan of operation (MPO) in 1977 improved the situation for 5-6 years with malaria cases reduced to about 2 million (Dash et al. 2008). The rising trend of malaria was facilitated by developments in various sectors which aimed to improve the national economy under successive 5-year plans. These developments led to increased irrigation, deforestation and rice-paddy and sugar cane cultivation, all of which profoundly influenced malaria transmission. With deteriorating malaria control, the government of India renewed approaches of combating the epidemic. Thus, malaria in India has been stratified into five ecotypes: tribal, rural with and without irrigation, urban, industrial and migration malaria (Sharma 1995).

The MPO, which is still in operation focused on vector control with IRS but additionally mobilized community participation in the programme and incorporated research into its three-pronged strategy (government efforts, malaria research and community participation). Local volunteers were identified and trained to operate as Fever Treatment depots (FTD) and Drug Distribution Centres (DDC). At FTDs blood smears were taken and treatment provided to all fever cases that visited then while in DDCs only treatment was provided (Misra 1999).

Malaria control initiatives in India are coordinated through the directorate of national vector borne diseases programme (NVBDP). This is the national level technical nodal office equipped with technical experts in the field of Public Health, Entomology, Toxicology and parasitology aspects of malaria. The Directorate is responsible for framing technical guidelines and policies as to guide the states for implementation of programme strategies. The key malaria control strategies in India are: (1) Early case Detection and Prompt Treatment (EDPT); (2) Vector control; (3) Community Participation; (4) Environmental Management & Source Reduction Methods; and (5) Monitoring and Evaluation of the programme (NVBDP) (India 1976). These strategies are summarised below.

1. Early case Detection and Prompt Treatment (EDPT). This is the main strategy of malaria control. The radical treatment is necessary for all the cases of malaria to prevent transmission of malaria. Chloroquine is the main anti-malaria drug for uncomplicated malaria. Drug Distribution Centres (DDCs) and Fever Treatment Depots (FTDs) have been established in the rural areas for providing easy access to anti-malarial drugs to the community. Alternative drugs for chloroquine resistant malaria are recommended as per the drug policy of malaria.

2. Vector control

   (i) Chemical Control. This includes the use of Indoor Residual Spray (IRS) with insecticides recommended under the programme; use of chemical larvicides like Abate in potable water; aerosol space spray during day time and Malathion fogging during outbreaks.

   (ii) Biological Control which involves the use of larvivorous fish in ornamental tanks, fountains and the biocides.

   (iii) Personal Prophylactic measures that individuals and communities can take up including the use of mosquito repellent creams, liquids, coils, mats, screening of the houses with wire mesh, use of bed nets treated with insecticide and wearing clothes that cover maximum surface area of the body

3. Community Participation. This involves sensitizing and involving the community for detection of Anopheles breeding places and their elimination. It also calls for collaboration with NGO schemes and involving them in programme strategies and working with other agencies at the community.

4. Environmental Management & Source Reduction Methods aimed at source reduction such as filling of the mosquito breeding places, proper covering of stored water and channelization of breeding source.

5. Computerized Management Information System (CMIS), field visits by state by State National Programme Officers, Malaria Research Centres and other ICMR Institutes and feedback to states on field observations for correction actions.

India achieved great success with the use of IRS over the years and other vector control options such as ITNs have been proposed as effective where the incidence of malaria is low and the vector bites late into the night. The choice of selective vector control measures would necessarily depend on the cost effectiveness of a strategy at a particular time and place. Researchers have assessed the cost effectiveness of different vector control strategies, including ITNs and IRS (Misra 1999).

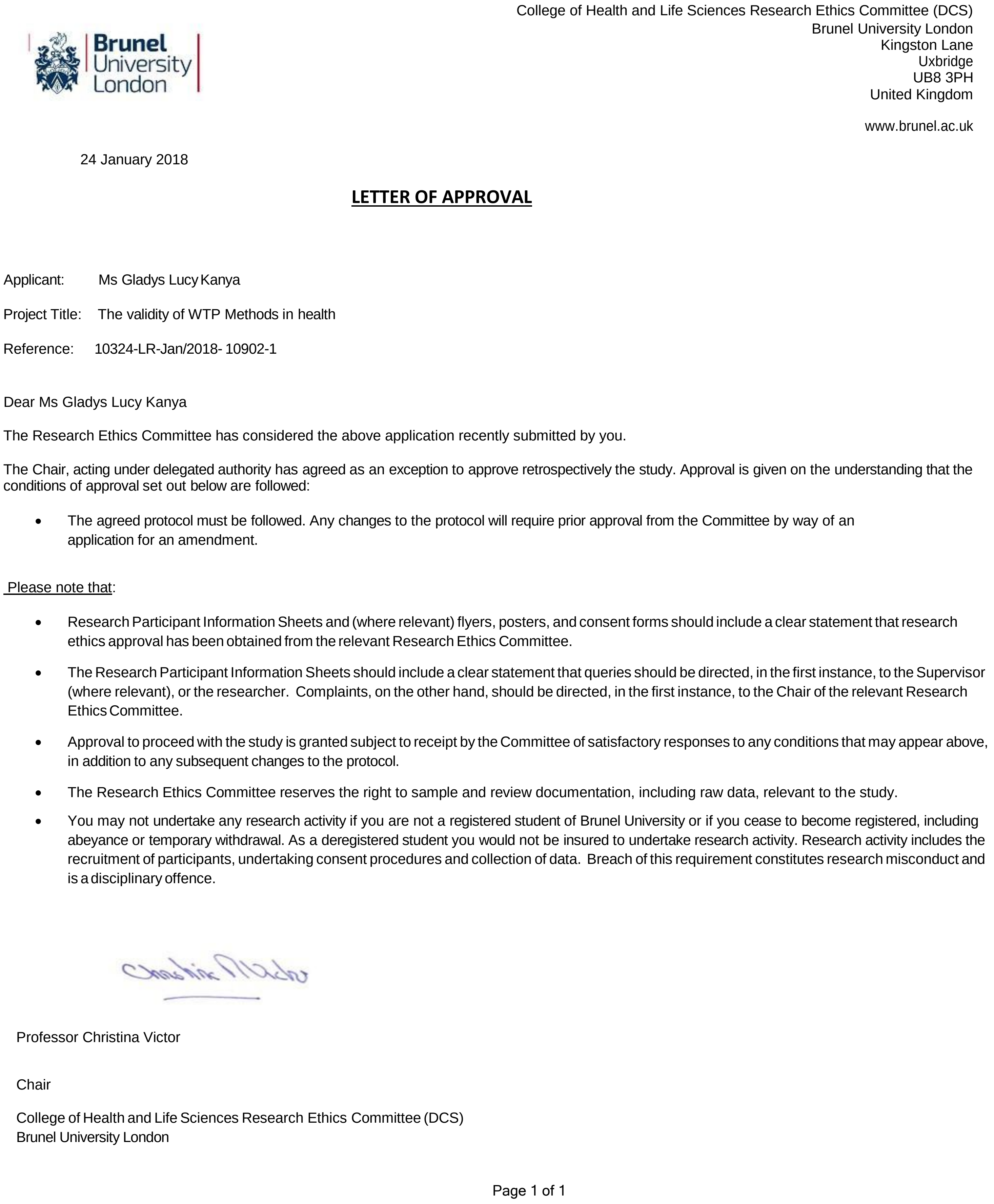
Appendix 20: Malaria WTP Study Hypothetical Survey Coding Frame

Malaria Control and Research Project: Health Economics Component Household Questionnaire for the WTP study

Variable Name	Variable in full	Categories
Intvill	Intervention group	4 = Outside trial area 1 = Treated mosquito nets 2 = Inhouse spray village 3 = Active case detection
hhsize	How many people live in this house	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 14, 15, 16, 23
Child<6	Children below the age of 6 years	Continuous
Sexme	Sex of the main earner	1 = Male 2 = Female
Sexr	Sex of the respondent	1 = Male 2 = Female
Edurc	Education respondent category	0 = Illiterate 1 = Primary 2 = Secondary 3 = Graduation
Edumec	Education main earner category	0 = Illiterate 1 = Primary 2 = Secondary 3 = Graduation
Occrc	Occupation resp cat	1 = Agriculture 2 = Animal husbandry 3 = Labour work 4 = Service 5 = Business 6 = Others
Oocmec	Main earner main occupation	1 = Agriculture 2 = Animal husbandry 3 = Labour work 4 = Service 5 = Business 6 = Others
Reli	Religion of the household	1 = Hindu 2 = Christian 3 = Muslim 4 = Parsee 5 = Others
Caste	Which caste does your household belong to	1 = Schedule caste 2 = Schedule tribe 3 = Other backward caste 4 = Other caste
Typehou	Type of house	1 = Kaccha 2 = Semi pucca 3 = Pucca
Mosqnus	Do you consider mosquito to be a nuisance	1 = Major nuisance 2 = Minor nuisance 3 = No nuisance
MosqmTotal	Total number of methods respondents know	Continuous variable
Premtotal	Prem1 and Prem2 combined	1 = Yes mosquito net

Variable Name	Variable in full	Categories
		0 = No mosquito net
Usetotal	Total number of methods respondents use	Continuous variable
Nown	How many do you own	1, 2, 3, 4, 5, 6, 7, 8,10, 11
Spytyr	This year was your house sprayed	1 = Sprayed completely 2 = Sprayed partially 3 = not sprayed
Spylyr	Was your house sprayed last year	1 = Sprayed completely 2 = Sprayed partially 3 = not sprayed
Mosqdistotal	Know of diseases that are caused by mosquito bites	1 = Malaria 0 = No malaria
lastyrmf	Members of the family suffering from Malaria last year	0, 1, 2, 3, 4, 5, 6, 7, 8, 10
lastmmf	Members of the family suffering from Malaria last month	1, 2, 3, 4, 6, 8
Exptremf	Expenditure incurred on treatment	0, 2, 4, 5, 6, 7, 10, 15, 18, 20, 25, 30, 35, 40, 45, 50, 55, 60, 70, 75, 80, 90, 100, 105, 120, 122, 125, 130, 135, 137, 138, 140, 150, 160, 170, 180, 185, 200, 210, 230, 240, 250, 260, 300, 350, 360, 400, 439, 500, 550, 570, 600, 650, 700, 800, 1000, 1100, 1200, 1400, 1500, 1700, 1800, 2200, 2250, 3190
TotalDisIncome	Total income (salary+ wages+ rent + interest + business+ remit + agriculture+ animals + fishers+ others)	Continuous variable
TotalExpY	Total yearly expenditure	Continuous variable
Wtbnets1	Willing to buy treated mosquito nets – including people who said yes to instalments	0 = No 1 = yes
Wtpbid1	Whether accept first bid	1=yes 0=no
Wtb2bidy	Willing to accept second higher bid	1=yes 0=no
Wtb2bidn	Willing to accept second lower bid	1=yes 0=no

Appendix 21: Brunel University Ethical Approval Letter



College of Health and Life Sciences Research Ethics Committee (DCS)
Brunel University London
Kingston Lane
Uxbridge
UB8 3PH
United Kingdom

www.brunel.ac.uk

24 January 2018

LETTER OF APPROVAL

Applicant: Ms Gladys Lucy Kanya

Project Title: The validity of WTP Methods in health

Reference: 10324-LR-Jan/2018- 10902-1

Dear Ms Gladys Lucy Kanya

The Research Ethics Committee has considered the above application recently submitted by you.

The Chair, acting under delegated authority has agreed as an exception to approve retrospectively the study. Approval is given on the understanding that the conditions of approval set out below are followed:

- The agreed protocol must be followed. Any changes to the protocol will require prior approval from the Committee by way of an application for an amendment.

Please note that:

- Research Participant Information Sheets and (where relevant) flyers, posters, and consent forms should include a clear statement that research ethics approval has been obtained from the relevant Research Ethics Committee.
- The Research Participant Information Sheets should include a clear statement that queries should be directed, in the first instance, to the Supervisor (where relevant), or the researcher. Complaints, on the other hand, should be directed, in the first instance, to the Chair of the relevant Research Ethics Committee.
- Approval to proceed with the study is granted subject to receipt by the Committee of satisfactory responses to any conditions that may appear above, in addition to any subsequent changes to the protocol.
- The Research Ethics Committee reserves the right to sample and review documentation, including raw data, relevant to the study.
- You may not undertake any research activity if you are not a registered student of Brunel University or if you cease to become registered, including abeyance or temporary withdrawal. As a deregistered student you would not be insured to undertake research activity. Research activity includes the recruitment of participants, undertaking consent procedures and collection of data. Breach of this requirement constitutes research misconduct and is a disciplinary offence.

Professor Christina Victor

Chair

College of Health and Life Sciences Research Ethics Committee (DCS)
Brunel University London

Page 1 of 1

Appendix 22: List of papers included in review of studies assessing WTP for TMNs

1. Aleme, A., Girma, E. & Fentahun, N. (2014). Willingness to pay for insecticide-treated nets in Berehet District, Amhara Region, Northern Ethiopia: implication of social marketing. *Ethiop J Health* Sci, 24(1), pp.75–84.

2. Biadgilign, S., Reda, A.A. & Kedir, H. (2015). Determinants of willingness to pay for the retreatment of insecticide treated mosquito nets in rural area of eastern Ethiopia. *International journal for equity in health*, 14(1), p.99.

3. Chase, C., Sicuri, E., Sacoor, C., Nhalungo, D., Nhacolo, A., Alonso, P.L. & Menendez, C. (2009). Determinants of household demand for bed nets in a rural area of Southern Mozambique. *Malaria Journal*, 8(1), p.132.

4. Gebresilassie, F. & Haile Mariam, D. (2000). Factors Influencing People's Willingness-to-buy Insecticide-treated Bednets in Arbaminch Zuria District, Southern Ethiopia. *J health popul nutr*. 29(3), pp.200–206.

5. Mujinja, P.M., Makwaya, C.K. & Sauerhborn, R., 2004. Gender and willingness to pay for insecticides treated bed nets in a poor rural area in Tanzania. *East African Medical Journal*, 81(12).

6. Onwujekwe, O., Chima, R., Shu, E., Nwagbo, D., Akpala, C. & Okonkwo, P. (2002). Altruistic willingness to pay in community-based sales of insecticide-treated nets exists in Nigeria. *Social Science and Medicine*, 54(4), pp.519–527.

7. 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. *Tropical Medicine & International Health*, 6(7), pp.545–553.

8. Onwujekwe, O., Hanson, K. & Fox-Rushby, J. (2004). Inequalities in purchase of mosquito nets and willingness to pay for insecticide-treated nets in Nigeria: challenges for malaria control interventions. *Malaria journal*, 3(6).

9. Onwujekwe, O. & Nwagbo, D. (2002). Investigating starting point bias: a survey of willingness to pay for insecticide-treated nets. *Social Science and Medicine*, 55(12). pp.2121–21230.

10. Taye, B. (2002). Willingness to Pay for Insecticide- Impregnated Bed Nets : The Case of Selected Rural Kebeles in Ilu Woreda of Western Shoa Zone. *Ethiopian Journal of Economics*, XI(1), pp.1–32.

Appendix 23: Household characteristics by intervention group

Variables	**Intervention group**				
	Treated Mosquito Net (TMN) (n=300)	**In-house Residual Spraying (IRS) (n=300)**	**Active Case detection (ACT) (n=300)**	**Outside trial area (OTA) (n=300)**	**Total Sample (n=1,200)**
Background characteristics					
Household size					
Mean (SD)	4.8 (1.8)	5.2 (2.3)	5.05 (2.6)	5.44 (2.2)	5.11 (2.26)
Range	1 - 14	1 - 15	1 -23	1 - 14	1 – 23
Children <6					
Mean (SD)	0.74 (0.93)	0.70 (0.99)	0.70 (0.92)	0.78 (1.01)	0.733 (0.97)
Range	0 – 4	0 – 5	0 – 4	0 – 5	0 - 5
Children 6+					
Mean (SD)	0.86 (1.01)	0.97 (1.17)	1.03 (1.21)	1.01 (1.08)	0.97 (1.12)
Range	0 – 4	0 – 8	0 – 7	0 - 5	0 - 8
Respondent Sex (%)					
Male	266 (88.7)	250	267	259	1,045 (86.83)
Female	34 (11.3)	50	33	41	158 (13.17)
chi2 = 5.39 Pr=0.145					
Main earner Sex (%)					
Male	283	280	280	271	1,122 (93.50)
Female	17	20	20	21	78 (6.50)
chi2 = 0.4936 Pr=0.920					
Respondent					
Main earner	277	265	276	249	1,067 (88.92)
No	23	35	35	51	133 (11.08)
Chi2=17.20 Pr=0.001					
Respondent education					
Illiterate	151	128	165	129	573
Primary	106	101	76	110	393
Secondary	36	65	54	51	206
Graduation	7	6	5	10	28
Chi2=24.22 Pr=0.004					
Res. Qualification 1					
Education	149	172	135	171	627
No Education	151	128	165	129	573
Chi2=12.94 Pr=0.005					

Variables	**Intervention group**				
	Treated Mosquito Net (TMN) (n=300)	**In-house Residual Spraying (IRS) (n=300)**	**Active Case detection (ACT) (n=300)**	**Outside trial area (OTA) (n=300)**	**Total Sample (n=1,200)**
Res. Qualification 2					
No education & Primary	257	229	241	239	966
Further education					
Chi2=8.55 Pr=0.036	43	71	59	61	234
Main earner education					
Illiterate	147	122	159	125	553
Primary	109	103	78	112	402
Secondary	38	67	58	53	216
Graduation	6	8	5	10	29
Chi2=24.20 Pr=0.004					
Main earner Qualification 1					
Education	153	178	141	175	647
No Education	147	122	159	125	553
Chi2=12.70 Pr=0.005					
Main earner Qualification 2					
No education & Primary	256	225	237	237	955
Further education					
Chi2=10.10 Pr=0.018	44	75	63	63	245
Respondent Occupation					
Agriculture	43	81	97	60	281
Animal Husbandry	22	23	23	13	81
Labour Work	168	125	119	133	545
Service	40	39	37	47	163
Business	10	18	13	27	78
Others	7	14	11	20	52
Chi2=51.47 Pr=0.000					
Main earner Occupation					
Agriculture	42	80	97	56	275
Animal Husbandry	20	19	21	13	73
Labour Work	161	123	117	135	536
Service	54	51	44	67	216
Business	20	21	14	28	83
Others	3	6	7	1	17
Chi2=52.11 Pr=0.000					

Variables	Intervention group				
	Treated Mosquito Net (TMN) (n=300)	**In-house Residual Spraying (IRS) (n=300)**	**Active Case detection (ACT) (n=300)**	**Outside trial area (OTA) (n=300)**	**Total Sample (n=1,200)**
Months in a year in main occupation					
Mean (SD)	10.77 (2.25)	10.90 (2.22)	10.50 (2.64)	10.54 (2.35)	10.68 (2.37)
Range	1 – 12	3 – 12	4 – 12	4 – 12	1 - 12
HH Religion					
Hindu	261	282	267	265	1,075
Christian	38	18	31	29	116
Muslim	0	0	2	6	8
Parsee	1	0	0	0	1
Others	-	-	-	-	-
Chi2=23.04 Pr=0.006					
Religion category					
Hindu	261	282	267	265	1,075
No Hindu	39	18	33	35	125
Chi2=9.0285 Pr=0.029					
Caste					
Scheduled caste	15	5	8	8	36
Schedule tribe	255	191	217	231	894
Other backward caste	19	73	66	41	199
Other caste	11	31	9	20	71
Chi2=69.38 Pr=0.000					
Type of house					
Kaccha	215	158	206	205	784
Semi pucca	52	77	51	36	216
Pucca	33	65	43	59	200
Chi2=39.04 Pr=0.000					
No. of rooms in house					
Mean (SD)	1.47 (0.83)	1.56 (0.84)	1.73 (1.21)	1.69 (0.91)	1.69 (0.91)
Range	1 - 8	1 – 6	1 - 15	1 – 6	1 - 6
Asset ownership					
LPG gas					
No	261	236	241	235	973

Variables	**Intervention group**				
	Treated Mosquito Net (TMN) (n=300)	**In-house Residual Spraying (IRS) (n=300)**	**Active Case detection (ACT) (n=300)**	**Outside trial area (OTA) (n=300)**	**Total Sample (n=1,200)**
Yes	39	64	59	65	227
Chi2=9.57 Pr=0.023					
Fan					
No	224	163	174	181	742
Yes	76	137	126	119	458
Chi2=30.24 Pr=0.000					
Radio					
No	206	213	207	191	817
Yes	94	87	93	109	383
Chi2=4.03 Pr=0.258					
Cupboard					
No	221	185	190	183	779
Yes	79	115	110	117	421
Chi2=13.82 Pr=0.003					
TV					
No	261	256	270	251	1,038
Yes	39	44	30	49	162
Chi2=5.6233 Pr=0.131					
Refrigerator					
No	286	277	282	273	1,118
Yes	14	23	18	27	82
Chi2=5.0788 Pr=0.166					
Truck					
No	296	286	289	290	1,161
Yes	4	14	11	10	39
Chi2=5.5920 Pr=0.133					
Car					
No	297	297	299	296	1,189
Yes	3	3	1	4	11
Chi2=1.743 Pr=0.627					
Bullock Cart					
No	263	258	245	257	1,023
Yes	37	42	55	43	177
Chi2=4.6324 Pr=0.201					

Variables	**Intervention group**				
	Treated Mosquito Net (TMN) (n=300)	**In-house Residual Spraying (IRS) (n=300)**	**Active Case detection (ACT) (n=300)**	**Outside trial area (OTA) (n=300)**	**Total Sample (n=1,200)**
Scooter					
No	276	265	261	261	1,063
Yes	24	35	39	39	137
Chi2=4.9687 Pr=0.174					
Moped					
No	291	285	293	90	1,159
Yes	9	15	7	10	41
Chi2=3.5102 Pr=0.319					
Cycle					
No	192	196	195	181	764
Yes	108	104	105	119	436
Chi2=2.0462 Pr=0.563					
Irrigated land in acres					
Mean (SD)	1.067 (4.77)	1.64 (3.850)	1.90 (5.07)	1.02 (2.44)	1.41 (4.177)
Range	0 – 55	0 – 37	0 – 46	0 – 23	0 - 55
Non-Irrigated land (Acres)					
Mean (SD)	0.95 (2.72)	0.72 (1.63)	0.82 (1.75)	0.84 (1.58)	0.8325 (1.979)
Range	0 – 37	0 – 15	0 – 15	0 - 12	0 – 37
No. of cows					
Mean (SD)	1.40 (2.08)	1.39 (1.88)	1.88 (2.42)	1.83 (2.62)	1.62 (2.28)
Range	0 – 12	0 – 11	0 – 14	0 - 19	0 - 19
No of Sheep					
Mean (SD)	0.43 (1.88)	0.28 (1.09)	0.32 (1.58)	0.38 (1.74)	0.355 (1.604)
Range	0 – 20	0 – 10	0 – 15	0 - 21	0 - 21
No of Hens					
Mean (SD)	1.31 (2.76)	0.82 (1.87)	0.99 (2.13)	1.39 (3.1)	1.12 (2.52)
Range	0 – 20	0 – 10	0 – 15	0 - 25	0 - 25
Malaria prevention variables					
Mosquito nuisance					
Major nuisance	**172**	246	249	249	916
Minor nuisance	119	48	47	40	254
No nuisance	9	6	4	11	30
Chi2=88.08 Pr=0.000					
Mosquito months in the village					

Variables	Intervention group				
	Treated Mosquito Net (TMN) (n=300)	In-house Residual Spraying (IRS) (n=300)	Active Case detection (ACT) (n=300)	Outside trial area (OTA) (n=300)	Total Sample (n=1,200)
January	1	3	2	2	8
February	1	3	1	2	7
March	3	2	1	1	7
April	2	2	3	3	10
May	6	2	5	5	18
June	6	9	6	10	31
July	55	40	66	99	260
August	148	120	122	146	536
September	22	50	60	4	136
October	26	33	18	12	89
November	20	13	8	7	48
December	9	20	7	8	44
DK	1	1	1	1	6
Chi2=132.5270 Pr=0.000					
Knows Mosquito nets					
Yes	299	286	270	255	1,110
No	1	14	30	45	90
Chi2=52.7087 Pr=0.000					
Knows Sprays					
Yes	151	209	185	170	713
No	149	91	115	130	487
Chi2=26.4688 Pr=0.000					
Knows coils					
Yes	49	66	48	72	235
No	251	234	252	228	965
Chi2=9.2867 Pr=0.026					
Knows Mats					
Yes	16	26	14	32	88
No	284	274	286	268	1,112
Chi2= 10.5952 Pr=0.014					
Knows Smoke					
Yes	262	260	271	272	1,065
No	38	40	29	28	135
Chi2=3.7642 Pr=0.288					

Variables	**Intervention group**				
	Treated Mosquito Net (TMN) (n=300)	**In-house Residual Spraying (IRS) (n=300)**	**Active Case detection (ACT) (n=300)**	**Outside trial area (OTA) (n=300)**	**Total Sample (n=1,200)**
Knows Odomos					
Yes	10	7	3	10	30
No	290	293	297	290	1,170
Chi2=4.5128 Pr=0.211					
Applies Oil					
Yes	18	8	10	16	52
No	282	292	290	284	1,148
Chi2= 5.4677 Pr=0.141					
Knows Sheets					
Yes	235	219	245	229	928
No	65	81	55	71	272
Chi2=6.7698 Pr=0.080					
Knows Fans					
Yes	95	156	144	150	545
No	205	144	156	150	655
Chi2=31.4745 Pr=0.000					
Knows Other measures					
Yes	1	6	6	0	13
No	299	294	294	300	1,187
Chi2=9.5652 Pr=0.023					
Total number of methods known					
Mean (SD)	3.78 (1.19)	4.14 (1.20)	3.98 (1.08)	4.02 (1.39)	3.98 (1.22)
Range	1 – 8	1 – 8	1 – 8	1 - 8	1 - 8
First preferred method					
Mosquito net	262	178	178	154	772
Spray	3	21	5	10	39
Smoke	14	51	51	65	181
Fan	17	41	57	46	161
Others	4	9	9	25	47
Chi2=129.0946 Pr=0.000					
Second preferred method					
Mosquito net	33	38	50	36	157
Spray	17	57	21	41	136
Smoke	139	75	90	91	395

Variables	Intervention group				
	Treated Mosquito Net (TMN) (n=300)	**In-house Residual Spraying (IRS) (n=300)**	**Active Case detection (ACT) (n=300)**	**Outside trial area (OTA) (n=300)**	**Total Sample (n=1,200)**
Fan	53	71	70	60	254
Others	58	59	69	72	258
Chi2= 64.0107 Pr=0.000					
Preferred method includes					
Mosquito net	295	216	228	190	929
Yes	5	84	72	110	271
No					
Chi2=114.4857 Pr=0.000					
Uses Mosquito Net					
Yes	297	51	51	66	465
No	3	249	249	234	735
Chi2= 613.8881 Pr=0.000					
Uses Mosquito Coils					
Yes	3	11	3	19	36
No	297	289	297	281	1,164
Chi2= 20.1604 Pr=0.000					
Uses Mats					
Yes	1	5	4	5	15
No	299	295	296	295	1,185
Chi2= 2.9030 Pr=0.407					
Uses Smoke					
Yes	231	234	253	266	984
No	69	66	47	34	216
Chi2= 18.4734 Pr=0.000					
Uses Odomos					
Yes	2	0	0	3	5
No	298	300	300	297	1,195
Chi2= 5.4226 Pr=0.143					
Applies oil to body					
Yes	5	3	3	11	22
No	295	297	297	289	1,178
Chi2=7.9642 Pr=0.047					
Uses Sheet to cover body					
Yes	229	216	241	243	929

Variables	Intervention group				
	Treated Mosquito Net (TMN) (n=300)	**In-house Residual Spraying (IRS) (n=300)**	**Active Case detection (ACT) (n=300)**	**Outside trial area (OTA) (n=300)**	**Total Sample (n=1,200)**
No	71	84	59	57	271
Chi2=8.8990 Pr=0.031					
Uses Fan					
Yes	66	132	126	108	432
No	234	168	174	192	768
Chi2=38.5417 Pr=0.000					
Uses Other methods					
Yes	0	2	0	3	5
No	300	298	300	297	1,195
Chi2=5.4226 Pr=0.143					
Total no of methods used by respondent					
Mean (SD)	2.78 (0.68)	2.18 (0.80)	2.27 (0.71)	2.41 (0.90)	2.41 (0.81)
Range	1 - 5	1 - 5	1 - 5	0 - 7	1 - 7
Whether house uses coil (has bought coil)					
Yes	3	11	3	19	36
No	297	289	297	281	1,164
Chi2= 20.16 Pr= 0.000					
Expenditure on coil during season	0.21 (2.24)	0.99 (5.57)	0.37 (4.45)	3.23 (18.53)	1.20 (10.05)
Mean (SD)	0 – 30	0 - 50	0 - 72	0 - 250	0 -250
Range					
Expenditure on coil off season					
Mean (SD)	0	0.33 (2.59)	0	1.12 (8.18)	0.3625 (4.31)
Range		0 – 30		0 - 125	0 - 125
Total expenditure on coils per month					
Mean (SD)	0.217 (2.24)	1.32 (7.79)	0.37 (4.45)	4.35 (25.92)	1.565 (13.84)
Range	0 – 30	0 – 75	0 – 72	0 - 375	0 - 375
Whether house uses mat (has bought mat)					
Yes	1	5	4	5	15

Variables	Intervention group				
	Treated Mosquito Net (TMN) (n=300)	**In-house Residual Spraying (IRS) (n=300)**	**Active Case detection (ACT) (n=300)**	**Outside trial area (OTA) (n=300)**	**Total Sample (n=1,200)**
No *Chi2=2.9030 Pr=0.407*	299	295	296	295	1,185
Expenditure on mat during season Mean (SD) Range	0.217 (3.75) 0 – 65	0.78 (7.78) 0 – 120	0.55 (4.79) 0 – 50	0.69 (5.99) 0 - 80	0.561 (5.77) 0 - 120
Expenditure on mat off season Mean (SD) Range	0.21 (3.75) 0 – 65	0.35 (4.022) 0 – 60	0.233 (2.372) 0 – 30	0.32 (2.96) 0 - 40	0.28 (3.33) 0 - 65
Total expenditure on mat per month Mean (SD) Range	0.433 (7.50) 0 – 130	1.13 (11.5) 0 – 45	0.783 (7.00) 0 – 80	1.01 98.71) 1.02 0 - 120	0.84 (8.869) 0 - 180
Whether house uses Odomos (has bought Odomos) Yes No *Chi2=5.4226 Pr=0.413*	2 298	0 300	0 300	3 297	5 1,195
Expenditure on Odomos during season Mean (SD) Range	0.233 (2.88) 0 – 40	0	0	0.5 (6.12) 0 - 100	0.18 (3.38) 0 - 100
Expenditure on Odomos off season Mean (SD) Range	0.667 (1.15) 0 – 20	0	0	0.233 (3.10)0 - 50	0.075 (1.65) 0 - 50
Total expenditure on Odomos per month Mean (SD) Range	0.3 (3.86) 0 – 60	0	0	0.733 (9.11) 0 - 150	0.258 (4.95) 0 - 150
Whether house uses Other methods (has bought Other methods) Yes	0 300	2 298	0 300	3 297	5 1,195

Variables	**Intervention group**				
	Treated Mosquito Net (TMN) (n=300)	**In-house Residual Spraying (IRS) (n=300)**	**Active Case detection (ACT) (n=300)**	**Outside trial area (OTA) (n=300)**	**Total Sample (n=1,200)**
No *Chi2= 5.4226 Pr=0.143*					
Expenditure on Other methods during season					
Mean (SD)	0	0.45 (5.944)	0	1.1 (13)	0.3875 (7.15)
Range		0 – 95		1.2 0 - 120	0 -200
Expenditure on Other methods off season					
Mean (SD)	0	0	0	0.55 (6.5)	0.1375 (3.25)
Range				0 - 100	0 - 100
Total expenditure on Other methods per month					
Mean (SD)	0	0.45 (5.94)	0	1.65	1.65 (19.50)
Range		0 – 95			0 - 300
Annual Household Income and Expenditure					
Annual Income (Rs.)					
Mean (SD)	24, 923.57 (40,953.12)	35,595.53 (56,732.03)	37,424.01 (76,464.81)	25,121.36 (30,193,49)	30,766 (54,218.99)
Range	0 – 400,000	2,250 – 700,000	2,500 – 807,200	2,000 – 197,400	0 – 807,200
Annual Expenses (Rs.)					
Mean (SD)	24,030.65 (32,314.03)	37,254.7 (53,189)	36,639.49 (58,088.26)	28,138.45 (28,251.6)	31,515 (45,145)
Range	3466-347,210	4,280 – 612, 532	2,890 – 551,000	2,565 – 218,310	2,565 – 612,532
Net Ownership, willingness to buy and willingness to pay					
Net Ownership					
Yes (%)	299 (99.67)	62 (20.67) 238 (79.33)	59 (19.67) 241	79 (26.33) 221	
No (%)	1 (0.33)		(80.33)	(73.67)	
Number of nets owned					
Mean (SD)	**2.9 (1.3)**	0.35 (0.85)	0.31 (0.73)	0.52 (1.08)	1.03 (1.51)
Range	**0 - 11**	0 - 5	0 - 4	0 - 6	1.04 0 - 11
Number of nets purchased through the market					
Mean (SD)	0.17 (0.54)	0.29 (0.80)	0.27 (0.71)	0.51 (1.07)	0.31 (0.81)
Range	0 - 4	0 - 5	0 - 4	0 - 6	0 - 6

Variables	**Intervention group**				
	Treated Mosquito Net (TMN) (n=300)	**In-house Residual Spraying (IRS) (n=300)**	**Active Case detection (ACT) (n=300)**	**Outside trial area (OTA) (n=300)**	**Total Sample (n=1,200)**
Number of nets distributed through the project Mean (SD) Range	2.74 (1.17) 0 – 8	0.02 (0.25) 0 - 3	0.01 (0.17) 0 – 3	**0 (0)**	0.69 (1.3) 0 - 8
Number of nets needed (*for those who did not have any)* Mean (SD) Range	2 (2) 2	2.47 (1.25) 0 – 7	2.52 (1.36) 0 – 9	2.21 (1.20) 0 – 6	**2.41 (1.28)** **0 - 9**
Number of additional nets needed (*for those who had some already)* Mean (SD) Range	0.22 (0.54) 0 – 3	1.87 (1.53) 0 – 7	1.91 (1.72) 0 – 9	2.01 (1.63) 2.02 0 - 6	0.911 (1.37) 0 - 9
Willingness to buy nets (Cash or Instalment) Yes (%) No (%)	159 (53) 141 (47)	263 (87.67) 37 (12.33)	263 (87.67) 37 (12.33)	258 (86) 42 (14)	943 (78.58) 257 (21.42)
Willingness to pay (Bid 1) Yes (%) No (%)	113 (37.92) 185 (62.08)	166 (55.70) 132 (44.30)	164 (55.59) 131 (44.41)	161 (54.03) 137 (45.97)	604 (50.80) 585 (49.20)
Willingness to pay (Bid 2 – Yes Bid1) Yes (%) No (%)	8 (2.68) 290 (97.32)	21 (7.05) 277 (92.95)	18 (6.10) 277 (93.90)	25 (8.39) 273 (91.61)	72 (6.06) 1,117 (93.94)
Willingness to pay (Bid 2 – No Bid1) Yes (%) No (%)	147 (49.33) 151 (50.67)	231 (77.52) 67 (22.48)	237 (80.34) 58 (19.66)	226 (75.84) 72 (24.16)	841 (70.73) 348 (29.27)
Maximum WTP Mean (SD) Range	38.28 (40.69) 0 – 200	64.19 (37.55) 0 – 300	64.22 (33.81) 0 – 220	62.06 (36.17) 0 - 200	57 (38) 0 - 300

Appendix 24: Willingness to buy nets _ Univariate and Base model estimation output

Dependent Variable 1: Willingness to buy nets		Univariate regression	Multiple regression (Base)
Variable (Base)	**Categories**	**Coefficient (Robust standard error)**	**Coefficient (Robust standard error)**
Interview village (Treated mosquito nets)	Active Case Detection	1.081***(0.224)	1.020***(0.332)
	In-house spray village	1.087***(0.213)	1.154***(0.343)
	Outside trial area	1.009***(0.228)	0.953***(0.363)
No. of people living in the house		0.0636***(0.0229)	0.0643*(0.0334)
No. of children below the age of 6 years		0.0521(0.0366)	-0.0554(0.0495)
Whether respondent main earner or not (Yes)		-0.0390(0.149)	0.0771(0.167)
Sex of the main earner (Male)		-0.305**(0.142)	-0.247(0.164)
Main earner's education (Graduation)	Illiterate	-0.346(0.345)	-0.681**(0.302)
	Primary	-0.264(0.350)	-0.552*(0.307)
	Secondary	-0.304(0.317)	-0.607*(0.311)
Respondent's Main Occupation (Agriculture)	Animal Husbandry	0.0751(0.214)	0.126(0.213)
	Business	0.0909(0.181)	0.288(0.202)
	Labour Work	-0.173(0.125)	-0.167(0.140)
	Others	-0.253(0.227)	-0.336(0.267)
	Service	0.00501(0.154)	0.00994(0.173)
Religion of the household (Christian)	Hindu	-0.166(0.302)	-0.106(0.270)
	Muslim	-0.620(0.459)	-0.0300(0.431)
	Parsee	-	-
Caste household belongs to (Scheduled caste)	Other backward caste	0.314(0.289)	0.343(0.264)
	Scheduled tribe	0.324(0.263)	0.735***(0.266)
Type of house (Kaccha)	Pucca	-0.0732(0.162)	-0.0510(0.171)
	Semi pucca	-0.388***(0.139)	-0.434***(0.125)
Total household disposable income (log)		0.0118(0.0138)	0.00604(0.0134)
Whether respondent considers mosquitoes to be a nuisance (Major nuisance)	Minor nuisance	-0.411***(0.119)	-0.0200 (0.122)
	No Nuisance	-0.912***(0.313)	-0.769***(0.267)
Total no. of mosquito measures known		0.0704*(0.0401)	0.0146 (0.0504)

Dependent Variable 1: Willingness to buy nets		Univariate regression	Multiple regression (Base)
Variable (Base)	**Categories**	**Coefficient (Robust standard error)**	**Coefficient (Robust standard error)**
Mosquito net among preferred methods (Yes))		-0.218(0.146)	0.178 (0.133)
Total no. of prevention methods used		-0.0807(0.0605)	0.131*(0.0743)
Whether household owns nets: Yes		-0.808***(0.151)	-0.218 (0.198)
No. of nets owned		-0.229***(0.0477)	-0.0328 (0.0688)
No. of nets purchased from market		-0.0282(0.0615)	-0.135(0.118)
If any disease is caused by mosquito bites (No)		0.262(0.247)	0.243(0.302)
No. of family members suffering from malaria last month		0.225(0.139)	-0.0315(0.175)
Expenditure incurred on malaria treatment (log)		0.120***(0.0307)	0.0973*(0.0509)
Constant		0.0673(0.194)	-0.356(0.770)
Pseudo R^2			0.1912
Linktest			0.07869[a]
Goodness of fit			6.45[b]
Observations			1,189

# The estimated parameters and asterisks show significance level of 1% (***), 5% (**) and 10% (*)

*** p<0.01, ** p<0.05, * p<0.1 [a] p=0.396 [b] p=0.5974

Appendix 25: Selection model estimation output: Willingness to pay for TMN at first bid, given willingness to buy nets

Dependent Variable 1: Willingness to pay (bid 1) given willingness to buy nets		**WTP1BID**	**WTBNETS**
Variable (omitted category)	**Categories / definition**	**Coefficient[#] (Robust standard error)**	**Coefficient[#] (Robust standard error)**
Interview village (Treated mosquito nets)	Active Case Detection	-0.471*(0.254)	1.111***(0.263)
	In-house spray village	-0.448*(0.256)	1.249***(0.272)
	Outside trial area	-0.386(0.262)	1.026***(0.276)
No. of people living in the house		0.0381*(0.0228)	
No. of children below the age of 6 years		-0.00276(0.0479)	
Whether respondent main earner or not (Yes)		0.161(0.145)	0.0385(0.160)
Sex of the main earner (Male)		-0.00767(0.168)	-0.323*(0.174)
Main earner’s education (Graduation)	Illiterate	-0.236(0.318)	-0.586*(0.316)
	Primary	-0.217(0.309)	-0.421(0.305)
	Secondary	-0.0364(0.307)	-0.538*(0.306)
Respondent’s Main Occupation (Agriculture)	Animal Husbandry	-0.328*(0.177)	0.0585(0.198)
	Business	-0.582***(0.180)	0.214(0.207)
	Labour Work	-0.449***(0.120)	-0.195(0.127)
	Others	-0.0746(0.239)	-0.352(0.260)
	Service	-0.379**(0.149)	0.0163(0.165)
Religion of the household (Christian)	Hindu	0.0255(0.137)	-0.124(0.160)
	Muslim	0.136(0.547)	-0.144(0.531)
	Parsee	4.465***(0.421)	6.252***(0.415)
Caste household belongs to (Scheduled caste)	Other backward caste	-0.224(0.224)	0.268(0.224)
	Scheduled tribe	-0.0893(0.226)	0.659***(0.230)
Type of house (Kaccha)	Pucca	0.0194(0.142)	-0.0412(0.159)
	Semi pucca	0.176(0.118)	-0.363**(0.124)*
Total household disposable income (log)		0.00451(0.0101)	0.00499(0.0110)
Whether respondent considers mosquitoes to be a nuisance (Major nuisance)	Minor nuisance	0.154(0.113)	-0.0273(0.114)
	No Nuisance	-0.181(0.282)	-0.785***(0.257)
Total no. of mosquito measures known		-0.0599(0.0403)	0.0251(0.0424)
Whether Mosquito is a preferred measure (Yes))		0.133(0.0937)	
Total no. of prevention methods used by household		0.0812(0.0658)	0.144**(0.0680)
Whether household owns nets: Yes		0.0915(0.186)	-0.211(0.187)
No. of nets owned		-0.000774(0.0774)	0.00420(0.0662)
No. of nets purchased from market		0.0746(0.102)	-0.135(0.104)

Dependent Variable 1: Willingness to pay (bid 1) given willingness to buy nets		WTP1BID	WTBNETS
Variable (omitted category)	**Categories / definition**	**Coefficient# (Robust standard error)**	**Coefficient# (Robust standard error)**
If any disease is caused by mosquito bites (No)		0.419*(0.243)	0.258(0.254)
No. of family members suffering from malaria last month		0.0132(0.103)	-0.0291(0.145)
Expenditure incurred on malaria related treatment (log)		-0.0164(0.0313)	0.0908**(0.0440)
Constant		0.632(0.599)	-0.0495(0.621)
Rho			-0.1151 (0.633)
Observations		1,189	1,189

# The estimated parameters and asterisks show significance level of 1% (***), 5% (**) and 10% (*)

Appendix 26: Willingness to pay for one TMN _ Univariate and Base model estimation output

Willingness to pay for nets _ bid 1		Univariate regression	Multiple regression (Base)
Variable (Base)	**Categories**	**Coefficient (Robust standard error)**	**Coefficient (Robust standard error)**
Interview village (Treated mosquito nets)	Active Case Detection	-0.235*(0.131)	-0.0530(0.311)
	In-house spray village	-0.234(0.148)	0.000595(0.337)
	Outside trial area	-0.253*(0.145)	0.0290(0.319)
No. of people living in the house		0.0595***(0.0213)	0.0516*(0.0282)
No. of children below the age of 6 years		0.0312(0.0425)	-0.0223(0.0528)
Whether respondent main earner or not (Yes)		0.126(0.128)	0.206(0.162)
Sex of the main earner (Male)		-0.206(0.157)	-0.124(0.185)
Main earner’s education (Graduation)	Illiterate	-0.576*(0.305)	-0.431(0.342)
	Primary	-0.446(0.295)	-0.383(0.324)
	Secondary	-0.245(0.311)	-0.171(0.324)
Respondent’s Main Occupation (Agriculture)	Animal Husbandry	-0.391**(0.185)	-0.348*(0.198)
	Business	-0.517***(0.174)	-0.574***(0.180)
	Labour Work	-0.564***(0.108)	-0.546***(0.128)
	Others	-0.348*(0.203)	-0.227(0.236)
	Service	-0.267(0.163)	-0.425**(0.180)
Religion of the household (Christian)	Hindu	-0.0782(0.140)	0.00728(0.145)
	Muslim	-0.197(0.483)	0.0484(0.544)
	Parsee	-	-
Caste household belongs to (Scheduled caste)	Other backward caste	-0.186(0.210)	-0.100(0.219)
	Scheduled tribe	-0.254(0.204)	0.166(0.243)
Type of house (Kaccha)	Pucca	0.193(0.118)	0.00139(0.148)
	Semi pucca	0.00902(0.134)	0.0213(0.141)
Total household disposable income (log)		0.0136(0.00965)	0.00604(0.0134)
Whether respondent considers mosquitoes to be a nuisance (Major nuisance)	Minor nuisance	0.235*(0.127)	0.169(0.136)
	No Nuisance	-0.601**(0.284)	-0.531*(0.271)
Total no. of mosquito measures known		0.0263(0.0342)	-0.0608(0.0438)

Willingness to pay for nets _ bid 1		**Univariate regression**	**Multiple regression (Base)**
Variable (Base)	**Categories**	**Coefficient (Robust standard error)**	**Coefficient (Robust standard error)**
Mosquito net among preferred methods (Yes))		0.223**(0.106)	0.165(0.113)
Total no. of prevention methods used		0.205***(0.0546)	0.112(0.0712)
Whether household owns nets: Yes		0.365***(0.0943)	0.0361(0.192)
No. of nets owned		0.130***(0.0386)	0.0320(0.101)
No. of nets purchased from market		0.222***(0.0639)	0.0216(0.126)
If any disease is caused by mosquito bites (No)		0.694***(0.258)	0.692**(0.294)
No. of family members suffering from malaria last month		0.0201(0.0697)	0.0170(0.117)
Expenditure incurred on malaria treatment (log)		0.00691(0.0180)	0.00552(0.0314)
Constant		0.582***(0.119)	-0.208(0.629)
Pseudo R^2			0.0595
Linktest			0.02766[a]
Goodness of fit			11.02[b]
Observations		932	932

# The estimated parameters and asterisks show significance level of 1% (***), 5% (**) and 10% (*) [a] p=0.893 [b] p=0.2004

Appendix 27: Selection model estimation _ Willingness to pay for TMN at second bid, given first bid

Dependent Variable 1: Willingness to pay (bid 2) given bid 1		**WTP1BIDy**	**WTP1BID**
Variable (omitted category)	**Categories / definition**	**Coefficient$^{\#}$ (Robust standard error)**	**Coefficient$^{\#}$ (Robust standard error)**
Interview village (Treated mosquito nets)	Active Case Detection	0.107(0.452)	-0.0540(0.314)
	In-house spray village	0.0691(0.465)	-8.59e-06 (0.339)
	Outside trial area	0.170(0.461)	0.0271(0.326)
No. of people living in the house		-0.0152(0.0629)	0.0518*(0.0293)
No. of children below the age of 6 years		0.0704(0.0815)	-0.0222(0.0528)
Whether respondent main earner or not (Yes)		0.0932(0.289)	0.207(0.165)
Sex of the main earner (Male)		0.0301(0.362)	-0.124(0.182)
Main earner's education (Graduation)	Illiterate	-0.341(0.559)	-0.433(0.352)
	Primary	-0.129(0.472)	-0.385(0.332)
	Secondary	-0.0936(0.416)	-0.174(0.341)
Respondent's Main Occupation (Agriculture)	Animal Husbandry	0.146(0.453)	-0.347*(0.205)
	Business	0.581(0.682)	-0.574***(0.180)
	Labour Work	-0.0744(0.516)	-0.545***(0.129)
	Others	0.473(0.494)	-0.225(0.254)
	Service	0.362(0.526)	-0.425**(0.180)
Religion of the household (Christian)	Hindu	-0.308(0.204)	0.00467(0.170)
	Muslim	-6.023***(0.622)	0.0451(0.571)
	Parsee	-5.405***(0.581)	3.885***(0.504)
Caste household belongs to (Scheduled caste)	Other backward caste	-0.665**(0.278)	-0.101(0.223)
	Scheduled tribe	-0.390(0.403)	0.163(0.261)
Type of house (Kaccha)	Pucca	0.226(0.259)	0.00115(0.148)
	Semi pucca	0.0630(0.226)	0.0207(0.141)
Total household disposable income (log)		-0.00639(0.0182)	0.00678(0.0113)
Whether respondent considers mosquitoes to be a nuisance (Major nuisance)	Minor nuisance	0.0360(0.239)	0.169(0.136)
	No Nuisance	-4.659***(0.355)	-0.531*(0.271)
Total no. of mosquito measures known		-0.072(0.0712)	-0.0612(0.0465)
Whether Mosquito is a preferred measure (Yes))		0.054(0.260)	0.166(0.118)
Total no. of prevention methods used by household		0.222*(0.126)	0.111(0.0776)
Whether household owns nets: Yes		-0.0004(0.275)	0.035(0.195)
No. of nets owned		-0.163(0.123)	0.0318(0.101)
No. of nets purchased from market		0.326*(0.169)	0.0228(0.133)
If any disease is caused by mosquito bites (No)		4.784	0.692**(0.294)
No. of family members suffering from malaria last month		-0.0828(0.172)	0.0174(0.116)

Dependent Variable 1: Willingness to pay (bid 2) given bid 1		WTP1BIDy	WTP1BID
Variable (omitted category)	**Categories / definition**	**Coefficient# (Robust standard error)**	**Coefficient# (Robust standard error)**
Expenditure incurred on malaria related treatment (log)		-0.0121(0.0486)	0.00552(0.0313)
Constant		-5.762***(1.513)	-0.198(0.686)
Rho			0.104(1.943)
Observations		932	932

# The estimated parameters and asterisks show significance level of 1% (***), 5% (**) and 10% (*)

Appendix 28: Willingness to pay for one TMN at a second higher bid _ Univariate and Base model estimation output

Dependent Variable 3: Willingness to buy nets _ bid 2_Yes		Univariate regression	Multiple regression (Base)
Variable (Base)	**Categories**	**Coefficient (Robust standard error)**	**Coefficient (Robust standard error)**
Interview village (Treated mosquito nets)	Active Case Detection	0.158(0.182)	-0.0161(0.418)
	In-house spray village	0.234(0.206)	-0.0264(0.452)
	Outside trial area	0.341*(0.191)	0.0902(0.448)
No. of people living in the house		0.0323(0.0279)	-0.00305(0.0388)
No. of children below the age of 6 years		0.0148(0.0553)	0.0418(0.0652)
Whether respondent main earner or not (Yes)		0.00257(0.165)	0.142(0.238)
Sex of the main earner (Male)		-0.0237(0.259)	0.0988(0.283)
Main earner’s education (Graduation)	Illiterate	-0.968***(0.316)	-0.422(0.412)
	Primary	-0.682**(0.295)	-0.245(0.339)
	Secondary	-0.470(0.306)	-0.130(0.360)
Respondent’s Main Occupation (Agriculture)	Animal Husbandry	-0.217(0.276)	0.0791(0.274)
	Business	0.263(0.253)	0.383(0.274)
	Labour Work	-0.422**(0.180)	-0.159(0.174)
	Others	0.337(0.262)	0.369(0.372)
	Service	0.213(0.212)	0.237(0.252)
Religion of the household (Christian)	Hindu	-0.255(0.160)	-0.282(0.187)
	Muslim	-	-
	Parsee	-	-
Caste household belongs to (Scheduled caste)	Other backward caste	-0.717***(0.190)	-0.616**(0.245)
	Scheduled tribe	-0.747***(0.194)	-0.298(0.295)
Type of house (Kaccha)	Pucca	0.506***(0.154)	0.252(0.226)
	Semi pucca	0.0499(0.174)	0.0831(0.199)
Total household disposable income (log)		0.00928(0.0135)	-0.00210(0.0160)
Whether respondent considers mosquitoes to be a nuisance (Major nuisance)	Minor nuisance	0.104(0.152)	0.0620(0.180)
	No Nuisance	-	-
Total no. of mosquito measures known		0.0994**(0.0430)	-0.0932*(0.0556)

Dependent Variable 3: Willingness to buy nets _ bid 2_Yes		Univariate regression	Multiple regression (Base)
Variable (Base)	**Categories**	**Coefficient (Robust standard error)**	**Coefficient (Robust standard error)**
Mosquito net among preferred methods (Yes))		0.179(0.162)	0.130(0.176)
Total no. of prevention methods used		0.304***(0.0691)	0.236**(0.0923)
Whether household owns nets: Yes		0.382***(0.141)	-0.0214(0.232)
No. of nets owned		0.0847**(0.0413)	-0.158(0.120)
No. of nets purchased from market		0.316***(0.0684)	0.329**(0.161)
If any disease is caused by mosquito bites (No)		-	-
No. of family members suffering from malaria last month		-0.0539(0.104)	-0.0443(0.163)
Expenditure incurred on malaria treatment (log)		-0.0185(0.0263)	-0.0157(0.0457)
Constant		-1.636***(0.157)	-1.216(0.915)
Pseudo R^2			0.1321
Linktest			-0.22875[a]
Goodness of fit			19.57
Observations		932	932

# The estimated parameters and asterisks show significance level of 1% (***), 5% (**) and 10% (*)[a] p=0.170 [b] p=0.0121

Appendix 29: Willingness to pay for one TMN at a second lower bid _ Base model estimation output

Dependent Variable 3: Willingness to buy nets _ bid 2_No		**Univariate regression**	**Multiple regression (Base)**
Variable (Base)	**Categories**	**Coefficient (Robust standard error)**	**Coefficient (Robust standard error)**
Interview village (Treated mosquito nets)	Active Case Detection	-0.129(0.197)	-0.184(0.449)
	In-house spray village	-0.324(0.210)	-0.291(0.451)
	Outside trial area	-0.335*(0.200)	-0.243(0.457)
No. of people living in the house		-0.0230(0.0281)	-0.0413(0.0361)
No. of children below the age of 6 years		-0.0233(0.0603)	0.0360(0.0702)
Whether respondent main earner or not (Yes)		0.268(0.181)	0.253(0.216)
Sex of the main earner (Male)		0.512(0.460)	0.628(0.384)
Main earner's education (Graduation)	Illiterate	-0.117(0.380)	-0.0999(0.368)
	Primary	-0.133(0.352)	-0.0957(0.342)
	Secondary	-0.0632(0.392)	0.0290(0.355)
Respondent's Main Occupation (Agriculture)	Animal Husbandry	-0.0920(0.259)	-0.260(0.276)
	Business	-0.419(0.258)	-0.559**(0.266)
	Labour Work	-0.508***(0.139)	-0.744***(0.184)
	Others	-0.530*(0.309)	-0.697**(0.338)
	Service	-0.318(0.201)	-0.544***(0.203)
Religion of the household (Christian)	Hindu	-0.160(0.225)	-0.0892(0.232)
	Muslim	-1.196**(0.542)	-0.597(0.602)
	Parsee	-	-
Caste household belongs to (Scheduled caste)	Other backward caste	0.402*(0.239)	0.442(0.281)
	Scheduled tribe	0.444**(0.220)	0.792***(0.247)
Type of house (Kaccha)	Pucca	-0.0903(0.150)	-0.0516(0.185)
	Semi pucca		0.0818(0.179)
Total household disposable income (log)		0.0225(0.0165)	0.0181(0.0176)
Whether respondent considers mosquitoes to be a nuisance (Major nuisance)	Minor nuisance	0.219(0.176)	0.204(0.191)
	No Nuisance	-0.426(0.380)	-0.211(0.347)

Dependent Variable 3: Willingness to buy nets _ bid 2_No		**Univariate regression**	**Multiple regression (Base)**
Variable (Base)	**Categories**	**Coefficient (Robust standard error)**	**Coefficient (Robust standard error)**
Total no. of mosquito measures known		-0.0259(0.0501)	-0.0325(0.0662)
Mosquito net among preferred methods (Yes))		0.0481(0.142)	-0.0491(0.149)
Total no. of prevention methods used		0.103(0.0745)	0.0336(0.0957)
Whether household owns nets: Yes		0.339**(0.145)	0.369(0.263)
No. of nets owned		0.0778(0.0644)	-0.126(0.144)
No. of nets purchased from market		0.139(0.0941)	0.137(0.151)
If any disease is caused by mosquito bites (No)		0.709***(0.207)	0.781***(0.223)
No. of family members suffering from malaria last month		-0.0437(0.100)	-0.112(0.144)
Expenditure incurred on malaria treatment (log)		0.00466(0.0273)	0.0573(0.0414)
Constant		1.524***(0.156	0.0149(0.860)
Pseudo R^2			0.0982
Linktest			0.5651[a]
Goodness of fit			9.49[b]
Observations		932	932

# The estimated parameters and asterisks show significance level of 1% (***), 5% (**) and 10% (*) [a] p=0.033 [b] p=0.3048

Appendix 30: Maximum Willingness to pay for nets _ Univariate and Base model estimation output

Dependent Variable 5: Maximum Willingness to pay for one net		Univariate regression	Multiple regression (Base)
Variable (Base)	**Categories**	**Coefficient (Robust standard error)**	**Coefficient (Robust standard error)**
Interview village (Treated mosquito nets)	Active Case Detection	0.755(2.966)	3.523(5.561)
	In-house spray village	0.620(3.266)	2.741(5.719)
	Outside trial area	-0.429(3.284)	2.726(5.674)
No. of people living in the house		-0.202(0.432)	-0.910**(0.390)
No. of children below the age of 6 years		-1.364(0.973)	-0.231(0.807)
Whether respondent main earner or not (Yes)		-0.5330(3.020)	-1.226(2.681)
Sex of the main earner (Male)		-2.826(3.471)	-3.713(2.779)
Main earner’s education (Graduation)	Illiterate	-26.55***(8.927)	-6.659(6.893)
	Primary	-22.40**(8.731)	-6.260(6.752)
	Secondary	-15.52*(9.367)	-5.287(6.962)
Respondent’s Main Occupation (Agriculture)	Animal Husbandry	-5.251(5.121)	-0.201(4.542)
	Business	-5.186(3.461)	-3.433(3.056)
	Labour Work	-11.52***(2.325)	-1.970(2.164)
	Others	4.716(5.724)	3.828(4.341)
	Service	1.689(2.838)	1.576(2.659)
Religion of the household (Christian)	Hindu	-1.306(2.565)	1.322(1.902)
	Muslim	-10.16(11.19)	-8.521(8.537)
	Parsee	0.842(2.222)	-16.31**(7.624)
Caste household belongs to (Scheduled caste)	Other backward caste	-11.76*(6.594)	-5.908(5.312)
	Scheduled tribe	-18.09**(7.060)	-9.675*(5.436)
Type of house (Kaccha)	Pucca	10.10***(3.382)	-0.977(2.870)
	Semi pucca	0.658(2.767)	-0.988(1.909)
Total household disposable income (log)		0.484**(0.192)	0.158(0.162)
Whether respondent considers mosquitoes to be a nuisance (Major nuisance)	Minor nuisance	5.402**(2.671)	2.618(1.894)
	No Nuisance	-15.17***(5.431)	-5.231(4.716)

Dependent Variable 5: Maximum Willingness to pay for one net		**Univariate regression**	**Multiple regression (Base)**
Variable (Base)	**Categories**	**Coefficient (Robust standard error)**	**Coefficient (Robust standard error)**
Total no. of mosquito measures known		2.190***(0.833)	0.216(0.627)
Mosquito net among preferred methods (Yes))		3.023(2.488)	-0.0645(1.971)
Total no. of prevention methods used		5.540***(1.482)	0.461(1.436)
Whether household owns nets: Yes		8.586***(2.476)	1.931(2.972)
No. of nets owned		2.094***(0.789)	0.278(1.619)
No. of nets purchased from market		7.239***(1.397)	1.124(2.145)
If any disease is caused by mosquito bites (No)		15.88***(4.426)	4.752(4.071)
No. of family members suffering from malaria last month		-3.181**(1.599)	-4.249***(1.638)
Expenditure incurred on malaria treatment (log)		-0.357(0.490)	0.690(0.563)
Willingness to buy at first bid (Yes)		22.32***(2.097)	11.52***(1.796)
Willingness to buy at Second higher bid (Yes)		52.28***(6.193)	42.46***(5.695)
Willingness to buy at Second lower bid (Yes)		28.89***(1.833)	15.85***(2.066)
Constant		72.68***(2.298)	58.17***(10.55)
Pseudo R^2			0.1079
Linktest			0.1311[a]
Observations		932	932

# The estimated parameters and asterisks show significance level of 1% (***), 5% (**) and 10% (*) [a]p=0.2615

Appendix 31: Correlation matrix for final variables included in meta-regression

Variable	Environment sector	Other sector	Health sector	Sample type same	Mode same	Bidding both	Open ended both	Concurrent surveys	Pay duration	Between sample comparison	Survey setting
Environment sector	1.0000										
Other sector	-0.5311	1.0000									
Health sector	-0.6497	-0.2990	1.0000								
Sample type same	0.1589	0.1371	-0.3020	1.0000							
Mode same	-0.2582	0.1371	0.1678	0.1923	1.0000						
Bidding both	-0.3363	-0.1548	0.5177	0.0868	0.0868	1.0000					
Open ended both	0.4074	-0.2164	-0.2647	0.1214	0.1214	-0.1370	1.0000				
Concurrent surveys	0.0437	0.2793	-0.2998	0.2491	0.2491	-0.2812	-0.0017	1.0000			
Pay duration	-0.2582	0.1371	0.1678	0.1923	1.0000	0.0868	0.1214	0.2491	1.0000		
Between sample comparison	-0.2607	0.0949	0.2084	-0.3864	-0.3864	-0.1548	-0.2164	0.0985	-0.3864	1.0000	
Survey setting	0.3721	-0.7006	0.2095	-0.0961	-0.0961	0.1085	0.1516	-0.3857	-0.0961	-0.1194	1.0000

Appendix 32: Rule of thumb for interpreting the size of a correlation coefficient

Size of Correlation	**Interpretation**
.90 to 1.00 (−.90 to −1.00)	Very high positive (negative) correlation
.70 to .90 (−.70 to −.90)	High positive (negative) correlation
.50 to .70 (−.50 to −.70)	Moderate positive (negative) correlation
.30 to .50 (−.30 to −.50)	Low positive (negative) correlation
.00 to .30 (.00 to −.30)	Negligible correlation

Source: (Hinkle et al., 2003)

Appendix 33: Pairwise correlation matrix for independent variables included in the regression analysis

	Int.village	HH size	Child <6yrs	Main earner	M.earner sex	M.earner edu	Resp Occu	Religion	caste	House type	Mosq. Nuisance	Prev, methods known	ITN pref. method	Total methods used	Owns Net	No. nets owned	Purchased net	Know Mosq. disease	Malaria episode last mnth	Malaria treat. Expenses	HH Income
Int.village	1.000																				
HH size	0.092	1.000																			
Child <6yrs	0.011	0.507	1.000																		
Main earner	-0.088	-0.034	-0.066	1.000																	
M.earner sex	0.018	-0.181	-0.049	-0.037	1.000																
M.earner edu	0.039	0.139	-0.058	-0.074	-0.097	1.000															
Resp Occu	0.012	-0.088	-0.000	-0.183	0.011	0.002	1.000														
Religion	0.022	0.011	0.029	-0.017	-0.047	-0.033	0.012	1.000													
caste	0.050	0.070	-0.085	0.008	0.013	0.417	-0.185	0.021	1.000												
House type	0.017	0.098	-0.097	0.001	-0.020	0.432	-0.081	-0.099	0.492	1.000											
Mosq. Nuisance	-0.173	-0.006	0.002	0.067	0.016	0.075	-0.054	-0.026	0.088	0.137	1.000										
Prev, methods known	0.053	0.089	-0.037	0.025	-0.065	0.314	-0.002	-0.013	0.271	0.255	0.016	1.000									
ITN pref. method	-0.271	0.004	0.037	0.023	0.003	0.027	0.011	0.023	-0.044	0.027	0.020	0.103	1.000								
Total methods used	-0.135	0.104	0.026	-0.020	-0.033	0.199	-0.010	0.084	0.097	0.110	0.005	0.412	0.146	1.000							
Owns Net	-0.502	0.069	0.037	0.019	-0.010	0.155	0.059	0.085	0.008	0.068	0.195	0.062	0.306	0.518	1.000						
No. nets owned	-0.536	0.215	0.119	0.050	-0.028	0.134	0.033	0.105	0.002	0.060	0.166	0.069	0.282	0.440	0.805	1.000					
Purchased net	0.140	0.245	0.017	0.019	-0.009	0.302	-0.024	0.087	0.231	0.231	0.015	0.245	0.089	0.376	0.454	0.474	1.000				
Know Mosquito disease	-0.106	-0.012	-0.019	0.014	-0.030	0.048	0.036	0.030	0.002	0.036	-0.159	0.092	0.101	0.096	0.082	0.032	0.050	1.000			
Malaria Episode last mnth	0.146	0.121	0.090	-0.069	-0.011	0.077	0.008	-0.030	0.021	0.029	-0.043	0.029 -	0.012 -	-0.021	-0.090	-0.076	0.016	-0.000	1.000		
Malaria treat. Expenses	0.097	0.038	0.047	-0.067	-0.024	0.088	0.000	0.005	0.074	0.065	-0.052 -	-0.049	-0.047	-0.068	-0.080	-0.069	-0.013	-0.008	0.299	1.0000	
HH Income	-0.038	-0.019	0.001	-0.008	-0.026	0.069	-0.015	0.012	0.030	0.023	0.014	0.059	0.051	0.075	0.074	0.064	0.083	-0.009	-0.078	-0.066	1.000