Exploring Alternative Methodologies for Robust Inferences: Applications in Environmental and Health Economics

Sapna Kaul

Dissertation submitted to the Faculty of the Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

In

Economics

Kevin J. Boyle, Chair

George C. Davis

Nicolai V. Kuminoff

Klaus Moeltner

Christopher F. Parmeter

September 13, 2013

Blacksburg, Virginia

Keywords: Benefit Transfer, Meta-Analysis, Survey Elicitation Effects, Program Evaluation

Copyright 2013

Exploring Alternative Methodologies for Robust Inferences: Applications in Environmental and Health Economics

Sapna Kaul

ABSTRACT

Researchers often invoke strong assumptions in empirical analyses to identify significant statistical outcomes. Invoking assumptions that do not sufficiently reflect the occurrence of true phenomenon reduces the credibility of inferences. Literature suggests that the potential effects of assumptions on credibility of inferences can be mitigated by comparing and combining insights from alternative econometric models. I use this recommendation to conduct robustness checks of commonly used methods in environmental and health economics. The first chapter proposes a novel nonparametric regression model to draw credible insights from meta-analyses. Existing literature on benefit-transfer validity is examined as an application. Nonparametric regression is found to be a viable approach for drawing robust policy insights. The second chapter proposes an alternative structural and simulations based framework to understand elicitation effects in survey response data. This analysis explains the structural mechanisms in which response anomalies occur and is important for building credible insights from survey data. The last chapter uses methods in program evaluation to investigate the impacts of institutional child deliveries on long-term maternal health in the context of developing countries. The outcomes of this analysis indicate that institutional deliveries positively affect maternal health in lower socio-economic states. Based on the findings of my three chapters, I recommend that researchers should combine insights from alternative models to mitigate the scope of specification bias in empirical outcomes and inform policy about the potential uncertainty that arises in uncovering the truth using statistical methods.

DEDICATION

To my parents, Vijay Kaul and Dr. Jawahar L. Kaul, for loving me unconditionally and supporting me in all of my endeavors; and to all my teachers for guiding me in my quest for knowledge.

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my adviser, Dr. Kevin Boyle for mentoring me academically and professionally at Virginia Tech. He encouraged me to pursue my research interests and constantly challenged me to think outside the box. This dissertation would not be complete without his guidance, support and thought provoking feedback.

I would like to extend my gratitude to all my committee members. Dr. Davis introduced health economics to me and stimulated me to conduct thorough and in-depth analyses of research questions. Dr. Kuminoff’s methodical approach to research and teaching always inspired me and he provided constructive feedback on my research. Dr. Moeltner’s lectures on environmental economics helped me to galvanize my research ideas and his incisive comments on my research were extremely helpful. Dr. Parmeter strengthened my understanding of data analysis, provided critical feedback on my research and was always there to address and clear my doubts.

My Ph.D. journey would not have fructified without the encouragement, support and good wishes of my family and friends. I would like to thank my parents, parents-in-law and brother, Divesh for having the confidence in my abilities and supporting me incessantly. I would like to thank all my friends for their unwavering support. Thank you Ranju, Nabin, Suruchi, Saurabh and Kristen for being my family away from home. I would also like to thank Dr. Wen You for being a friend and mentor at Virginia Tech. Most importantly, I would like to thank my best friend and husband, Sahil for his love and patience. This dissertation would certainly not be complete without his encouragement, involvement and understanding.

TABLE OF CONTENTS

LIST OF FIGURES

Chapter II

Chapter III

Chapter IV

LIST OF TABLES

Chapter II

Chapter III

Chapter IV

Chapter I. Introduction

A fundamental policy question in empirical economics is whether we can uncover the true events with precision and certainty. This is critical because the underlying truth is always unknown and to identify it we make identifying assumptions regarding the data generating processes. If the data are representative of the phenomena we are trying to evaluate, then the certainty with which we can claim the credibility of inferences essentially depends on whether assumptions employed in the empirical analyses sufficiently reflect the occurrences of true events. However, researchers often make strong assumptions for identification of statistical outcomes that can reduce the credibility in conclusions (Manski, 2011 & 2010). Clearly, a tradeoff between identification and credibility exists in inference building and we need to explore potential research methodologies that assist in resolving this tradeoff.

Manski (2011) proposes a procedure for minimizing the effect of theoretical and statistical assumptions on inference building. He suggests that *"a researcher can resolve the tension between the credibility and power of assumptions by posing alternative assumptions of varying credibility and determining the conclusions that follow in each case"* (pp. 262). This is similar in flavor to Leamer's (1983 & 1985) remedy for "*taking the con out of econometrics*" where he proposes conducting sensitivity analysis of results by changing specifications and functional forms of empirical methodologies. These suggestions essentially indicate that researchers should investigate the robustness of their findings in different settings and the common findings can be confidently used to inform policy.

In this dissertation, I follow these recommendations to conduct robustness checks of commonly used methods in environmental and health economics. In two of my chapters, I use alternative econometric models to divulge robust, data-driven insights that have economic relevance. In the third chapter, I propose alternative theory-based simulations to investigate anomalies in empirical outcomes of survey research. While I investigate topics in environmental and health economics, the methods and procedures outlined in my dissertation can be easily employed in others fields of study. Below, I briefly outline the motivation, methods and findings of each chapter.

Chapter II. What Can We Learn From Benefit Transfer Errors? Evidence from 20 Years of Research on Convergent Validity

This chapter proposes procedures that help in drawing credible insights from meta-analysis, which is a research assessment tool used for summarizing results of existing empirical studies. A meta-regression investigates the relationship between study designs/characteristics and effect size (a common unit for comparing study outcomes). The standard practice is to employ parametric regression models (e.g. ordinary least squares, weighted least squares etc.) to estimate the meta-regression (Nelson and Kennedy, 2009). However, the parametric models invoke strong assumptions (e.g. linearity, separability etc.) that can potentially reduce the credibility of inferences. In this chapter, an alternative method for estimating meta-regression is explored that can potentially minimize the specification error that arises because of putting restrictions on the relationship between the regressand and regressors.

A novel nonparametric regression model is proposed to draw robust insights from meta-regression. In contrast to the least squares method, the nonparametric method computes joint conditional density of the dependent variable at each observation of the independent variables without putting any restriction on the functional form of meta-regression. The estimated density function allows us to capture the joint effect of regressors on the regressand. In parametric settings, we will have to incorporate several dummy variables to capture interactions effects that cause inefficiency in estimation due to loss of degrees of freedom. In addition, nonparametric regression provides a distribution of marginal effects that can be used to investigate the robustness of parametric point estimates.

The above framework is used to conduct a robust meta-analysis of existing benefit-transfer, convergent-validity studies. Benefit transfer is one of the most commonly used methods for conducting costs and benefits analyses at the U.S. Environmental Protection Agency and other environmental agencies around the world (U.S. EPA, 2010). Convergent-validity studies have investigated the accuracy of benefit transfers; however, there is little or no consensus on research designs that improve the performance of benefit transfers (Johnston and Rosenberger, 2010). To resolve this, I conduct a meta-analysis of over thirty convergent-validity studies using standard and nonparametric regression methods.

The collective findings of the weighted least squares and nonparametric regression models help us to draw important insights. This analysis confirms the stylized facts (e.g. function transfers perform better than value transfers and geographical similarity improves benefit transfer accuracy) from the literature and also proposes some new insights (e.g. quantity changes, use of multiple studies and the contingent-valuation methods tend to improve benefit transfer accuracy). These insights are important for conducting future benefit transfers that will be designed to assess the benefits/impacts of existing policy regulations. Additionally, the analysis shows that the nonparametric regression method is a viable approach for drawing robust insights in a meta-analysis and testing the robustness of the traditional least squares model. Chapter II has been published in the Journal of Environmental Economics and Management and the copyright is assigned to Elsevier Inc.

Chapter III. Elicitation Effects in Surveys: A Structural and Simulations Based Analysis

This chapter examines response anomalies in survey data. Surveys are commonly used to support policy-making and academic research. They are extensively used to elicit preferences for changes in environmental quality and health outcomes (Ryan and Watson, 2009; Hanley et al., 1998). Inferences based on survey data assume that survey responses are procedurally invariant to the framing of the questions. However, empirical research shows that framing of questions have unintended consequences for survey responses (Conti and Pudney, 2011; Hurd et al., 1998). Mechanisms (e.g. anchoring, yea-saying etc.) have been proposed that explain anomalies in response data. However, these mechanisms have been largely identified through ex-post story telling or ex-post econometric adjustments, not through the development of a fundamental understanding of the data-generation processes. Additionally, we lack a clear understanding on whether these mechanisms are different concepts or similar concepts with different names, or empirically distinguishable.

In this chapter, an alternative structural framework is proposed to understand how different mechanisms create anomalies in response data. This approach is similar to that of Bernheim and Rangel's (2009) approach where they propose a general model of choice that explains nonstandard behavioral models. To conduct this exploration, I focus on the contingent-valuation survey format that is commonly used to elicit preferences and willingness to pay for non-market goods/services (Carson, 2000). Specifically, I investigate the dichotomous choice question framing where respondents are asked if they are willing to pay a stated bid amount. Existing literature indicates that response data from dichotomous choice surveys can potentially be affected by mechanisms such as anchoring, yea-saying, warm glow etc. (Boyle, Johnson and McCollum, 1997; Blamey, Bennett and Morrison, 1999; Leggett et al., 2003).

A utility theoretic model is used to explain how individuals respond to dichotomous choice questions. Elicitation effects are incorporated in this model (via mechanisms suggested in the literature) to explain their influence on observed response data. Statistical simulations are used to investigate the empirical outcomes of the theoretical model, and to test whether the actual and observed cumulative density functions of response data differ significantly. Using this analysis, we are able to explain the structural mechanisms in which response anomalies occur. Some mechanisms are observationally distinguishable and some are empirically equivalent. The outcomes of simulations show that specific elicitation effects can explain the observed empirical cumulative density functions of response variables, and that some ex-post story telling is not relevant. This analysis is important for building credible insights for policy making via survey data and it aids in uncovering potential areas where more research is required.

Chapter IV. Does Location Matter? The Effect of Institutional Versus Home Based Deliveries on Maternal Health

In this chapter, I employ the treatment evaluation approach to investigate the relationship between access to maternal health services and maternal health outcomes. Maternal health includes the health of mothers during pregnancy, childbirth and postpartum periods. Using observational data I examine the treatment effects associated with services provided during childbirth in the context of developing countries. Estimation of treatment effects using observational data can potentially be affected by self-selection bias that arises because of the correlation between the treatment assignment, individual characteristics and health outcomes. To avoid this, I use matching that minimizes the selection bias by creating a similar control group and estimating the average difference between the outcomes of the treated and the counterfactual control group. Similarity can be address by matching on observed characteristics of participants; however, a large number of characteristics can slow down the estimation process because of the curse of dimensionality (Heinrich, Maffioli, Vazquez, 2010). To counter this, I use propensity scores (conditional probability of receiving a treatment) to match the treated and non-treated participants. Propensity scores are estimated using a logistic model and the nearest neighbor and caliper matching algorithms are used to create the control group.

To estimate the treatment effects I use the location of child delivery (home versus health institution) as a proxy for services received during childbirth. In developing countries, traditional birth attendants perform home-based deliveries using outdated procedures that are known to be harmful for mothers and children (Goodburn et al., 2000). Unskilled delivery assistance has several physical and psychological health consequences for mothers (Filippi et al., 2006). These consequences can potentially affect mothers' weight status and households incur financial losses in the process of treating the adverse health impacts, which is known to cause weight loss in women in developing countries (McLaren, 2007). Therefore, I examine whether location of delivery has any effect on mother's long-term body mass index, which is frequently used as an indicator of adult health status (NIH, 1998).

The treatment effects are estimated using household survey data for two spatially and demographically diverse states (Bihar and Gujarat) in India. Bihar is one of the poorest states in India whereas Gujarat is one of the fastest growing Indian states (IIPS and Macro International, 2008a, 2008b). Our propensity score matching results show that mothers in Bihar who use health institutions for child deliveries tend to be healthier than those who deliver at home. The wealthier state, Gujarat, where more advance home-delivery practices may be employed shows insignificant treatment effects. These findings appear to be robust to specifications of the regression model and matching algorithms. The results suggest that there is a need to reevaluate maternal health care services in Bihar and other similar lower-income status states in India and elsewhere. First, the quality of home-based deliveries should be improved in poorer states like Bihar. Second, policies (investment in maternal health infrastructure, providing health and nutrition based education etc.) should be targeted to increase accessibility of and promote demand for institutional deliveries in poor states.

Conclusions

More often than not, empirical research focusses on confidence intervals to imply uncertainty. However, this estimation uncertainty is different from the uncertainty in results that is caused by putting assumptions on the econometric models. Researchers need to go beyond reporting confidence intervals and conduct robustness checks of their findings with respect to the modeling assumptions. Based on the findings of my three chapters, I recommend that empirical research should combine insights from alternative models that vary in the strength of implied assumptions. This helps in two ways. First, instead of a point estimate we can provide a range of effects that explains the uncertainty caused by identifying assumptions and also helps in informing policy about the potential distribution of impact effects. Second, more confidence can be placed on conclusions that follow from each of the methodologies whereas conclusions that are specific to the set of maintained assumptions are less likely to be credible. Using these procedures, we can attain deeper understanding of the credibility of our conclusions, which is critical for informing policy in any field of study.

References

Bernheim B. D. and A. Rangel. 2009. Beyond Revealed Preference: Choice-Theoretic Foundations for Behavioral Welfare Economics. The Quarterly Journal of Economics, 124(1): 51 – 104.

Blamey, R.K., J.W. Bennett and M.D. Morrison. 1999. Yea-saying in Contingent Valuation Surveys. Land Economics, 75 (1): 126-141.

Boyle, Kevin J., F. Reed Johnson, and Daniel W. McCollum. 1997. Anchoring and Adjustment in Single-Bounded, Contingent-Valuation Questions. American Journal of Agricultural Economics, 79(5): 1495-500.

Carson, Richard T. 2000. Contingent Valuation: A User's Guide. Environmental Science and Technology, 34(8): 1413-1418.

Conti, G., and S. Pudney. 2011. Survey design and the Analysis of Satisfaction. The Review of Economics and Statistics, 93 (3): 1087-1093.

Filippi, V., Ronsmans, C., Campbell, O., Graham, W. J. , Mills, A., Borghi, J., Koblinsky, M., Osrin, D., 2006. Maternal Health in Poor Countries: The Broader Context and a Call for Action. The Lancet, 368: 1535-1541.

Goodburn EA, Chowdhury M, Gazi R, Marshall T, Graham W. 2000. Training Traditional Birth Attendants in Clean Delivery Does Not Prevent Postpartum Infection. Health Policy and Planning, 15(4): 394-399.

Heinrich C., A. Maffioli, G. Vazquez. 2010. A Primer for Applying Propensity Score Matching, Impact – Evaluation Guidelines. Inter – American Development Bank.

Hurd, M.D., D. McFadden, H. Chand, L. Gan, A. Menill and M. Roberts. 1998. Consumption and Savings Balances of the Elderly: Experimental Evidence on Survey Response Bias. Chapter 8 in Frontiers in the Economics of Aging, D.A. Wise (ed.). University of Chicago Press.

Johnston, R.J., R.S. Rosenberger. 2010. Methods, Trends and Controversies in Contemporary Benefit Transfer. Journal of Economic Surveys, 24 (3): 479-510.

Hanley, N., D. MacMillan, R.E. Wright, C. Bullock, I. Simpson, D. Parsisson and B. Crabtree. 1998. Contingent Valuation Versus Choice Experiments: Estimating the benefits of Environmental Sensitive Areas in Scotland. Journal of Agricultural Economics, 49 (1): 1-15.

Leamer, E. 1985. Sensitivity Analyses Would Help. American Economic Review, 75(3): 308–313.

Leamer, E. 1983. Let’s Take the Con Out of Econometrics. American Economic Review, 73(1): 31–43.

Leggett, C.G., N.S. Kleckner, K.J. Boyle, J.W. Duffield and R.C. Cameron. 2003. Social Desirability Bias in Contingent Valuation Surveys Administered Through In-person

Interviews. Land Economics, 79 (4): 561-575.

Manski C.F. 2011. Policy Analysis with Incredible Certitude. The Economic Journal, 121: F261 – F289.

Manski, C.F. 2007. Identification for Prediction and Decision. Harvard University Press, Cambridge.

National Institute of Health, National Heart, Lung and Blood Institute. Clinical Guidelines on the Identification, Evaluation and Treatment of Overweight and Obesity in Adults: The Evidence Report. Obesity Research, 1998.

Nelson, J.P., P.E. Kennedy. 2009. The Use (and Abuse) of Meta-Analysis in Environmental And Natural Resource Economics: An Assessment. Environmental and Resource Economics 42(3): 345-377.

Ryan, M., and V. Watson. 2009. Comparing Welfare Estimates from Payment Card Contingent Valuation and Discrete Choice Experiments. Health Economics, 18 (4): 389-401.

U.S. Environmental Protection Agency. 2010. Guidelines for Preparing Economic Analyses. EPA 240-R_10-001 (prepublication edition).

http://yosemite.epa.gov/ee/epa/eed.nsf/pages/Guidelines.html/$file/Guidelines.pdf, accessed December 31, 2010.

Chapter II. What Can We Learn from Benefit Transfer Errors? Evidence from 20 Years of Research on Convergent Validity

Abstract

A nonparametric approach to meta-analysis is used to identify modeling decisions that affect benefit transfer errors. The meta-data describe the results from 31 empirical studies testing the convergent validity of benefit transfers. They evaluated numerous methodological procedures, collectively reporting 1071 transfer errors. Our meta-regressions identify several important findings, including: (1) the median absolute error is 39%; (2) function transfers outperform value transfers; (3) transfers describing environmental *quantity* generate lower transfer errors than transfers describing *quality* changes; (4) geographic site similarity is important for value transfers; (5) contingent valuation generates lower transfer errors than other valuation methods; and (6) combining data from multiple studies tends to reduce transfer errors.

Key Words: benefit transfer, function transfer, convergent validity, meta-analysis

1. Introduction

Benefit transfer is one of the most common methods for conducting benefit-cost analysis at the U.S. Environmental Protection Agency (U.S. EPA, 2010). When it is too time-consuming or expensive to directly estimate the monetary benefits of a policy, a surrogate measure is produced from preexisting estimates. A benefit transfer takes preexisting values from a study case (or cases) to develop a customized benefit estimate for a new policy case.[1] A "value transfer" simply substitutes a point estimate from a previous study or, in some cases, the mean or median of several point estimates. A "function transfer" predicts benefits using a previously calibrated function describing how values vary with the characteristics of people and places (Loomis and Rosenberger, 2006).[2]

Since a 1992 issue of *Water Resources Research* (Vol. 28, 3) raised academic interest in benefit transfers among environmental economists, at least 40 studies have investigated the empirical accuracy of this method using tests of convergent validity, which considers the difference between two different estimates of the same theoretical construct. These tests are designed to measure benefit transfer error, defined as the difference between a benefit measure estimated using original data (i.e. the policy case) and a surrogate for that benefit measure based on preexisting estimates (i.e. the study cases). Two stylized facts have emerged. Function transfers generate lower transfer errors than value transfers, and greater similarity between the study and policy cases reduces transfer errors (Boyle et al., 2010; Johnston and Rosenberger, 2010; Kirchhoff, Colby and LaFrance, 1997; Rosenberger and Phipps, 2007). The apparent lack

[1] We adopt the phrases *study case* and *policy case* from the U.S. EPA's 2010 *Guidelines for Preparing Economic Analyses* (U.S. EPA, 2010). This is a change from the conventional terminology, *study site* and *policy site*. Benefits are not always transferred to new geographic sites. Some transfers occur at the same physical location, using past values to assess current situations or predict future outcomes. Thus, benefits may be transferred to a new policy case at the same study site or to a new policy case at a different site.

[2] Readers seeking background on concepts, terminology, and methods in the benefit transfer literature are directed to the surveys prepared by Bergstrom and Taylor (2006), Boyle et al. (2010), and Johnston and Rosenberger (2010).

of consensus on other methodological features of the transfer process has made it difficult to define specific protocols for the conduct of benefit transfers and to develop a clear agenda for research. Johnston and Rosenberger (2010) observe, the "*complexity and relative disorganization of the (academic) literature may represent an obstacle to the use of updated (benefit-transfer) methods by practitioners.*"

The purpose of this paper is to investigate if specific modeling decisions can be identified that affect benefit transfer errors, and this is done using a new statistical approach to meta-analysis. Our quantitative research design complements and extends previous qualitative assessments of the literature (Rosenberger and Phipps, 2007; Rosenberger and Stanley, 2006). We begin by systematically reviewing empirical studies on the convergent validity of benefit transfers conducted over the past 20 years. These studies tested a tremendous variety of methodological procedures, collectively reporting *more than one thousand* benefit transfer errors. Most of these observations come from studies conducted in the United States and Western Europe. The applications cover a wide variety of amenities. Examples include access to forest, park, and lake recreation; hunting; changes in the quantity and quality of water in lakes, rivers, and coastal areas; air quality; exposure to ultraviolet radiation; freshwater fishing in streams and rivers; proximity to various types of open space; farmland amenities; and measures of the overall ecological health of watersheds, wetlands, and rivers.

It is standard practice in meta-analyses to use linear meta-regressions with robust standard errors to distill the collective findings on important questions in the field of environmental economics (Nelson and Kennedy, 2009). However, we have some concerns about the credibility of a *linear* meta-regression in the context of our analysis. The modeling decisions that comprise a benefit transfer are represented by binary variables that can interact in complex ways to influence benefit transfer errors (Boyle et al., 2009). Any specific variable can have a unique effect on the outcome being investigated (meta-equation regressand) when combined with sets of other regressors. For example, similarity between study and policy cases may be crucial to the accuracy of a value transfer, but not as important for a function transfer that can be calibrated to policy case conditions. A statistical approach that allows for these interaction effects has the potential to provide richer insights about the literature being investigated. The typical linear meta-regression imposes separability of each of the regressors. Capturing all potential interaction effects would require adding an intractable number of regressors to the meta-equation. Therefore, we propose a new nonparametric approach to meta-analysis that does not impose the linearity and separability assumptions. We contrast the insights from our new approach with a conventional parametric approach.

Nonparametric meta-analysis is particularly useful for our application to benefit transfer because it avoids the need to impose a-priori restrictions on the functional relationship between benefit-transfer errors and the various modeling decisions that comprise the transfer process. The nonparametric approach generates response effects for the combination of regressors at every data point, which yields ranges of effects instead of the parametric point estimates of average impacts. This allows us to recognize that there are multiple possible effects for each transfer characteristic, with specific impacts depending on the empirical context defined by other methodological choices made by analysts in implementing a benefit transfer.

We follow Charles Manski's (2007) "bottom-up" approach to data analysis where we start with the less restrictive nonparametric analysis and then move to a more restrictive parametric analysis. We use the nonparametric model to estimate the ranges of impacts for the benefit transfer characteristics that can be identified from variation in the meta-data. Then we add the linearity and separability restrictions that are commonly used in meta-regressions and repeat the estimation using weighted least squares. Following Manski's logic, we recognize that the credibility of inference based on our results is decreasing in the strength of the parametric restrictions that we impose on the meta-regression as findings may not be robust to relaxation of the imposed assumptions in the parametric analysis.

Our parametric and nonparametric models allow us to distill several important findings from the literature. First, our analysis confirms the stylized fact that function transfers outperform value transfers. Second, benefit transfers valuing changes in environmental *quality* (e.g. an increase in river clarity) almost always have larger errors than transfers describing *quantity* changes (e.g. an increase in the quantity of domestic water availability). Third, the geographic proximity of the study and policy locations tends to reduce transfer errors, especially for value transfers. Fourth, drawing on information from multiple preexisting studies (as opposed to a single study) also tends to reduce transfer errors, especially for function transfers. Lastly, for the stated-preference applications, contingent valuation generates lower transfer errors than choice modeling. For revealed preference applications, site demand models perform better than site choice models. Meta-analysis generates transfer errors that lie in between the stated and revealed preference models.

2. Conceptual Framework

The process of benefit transfer begins by defining the relevant measure of benefits. Consider a public policy that is expected to change the quality of an environmental amenity from q^0 to q^1 at policy case j. A partial equilibrium Hicksian measure of willingness to pay (wtp) for this change is defined as:

$$V_{ij}\left(p_j, x_j, q_j^1, y_i - wtp_{ij};\ \alpha_i, d_i\right) = V_{ij}\left(p_j, x_j, q_j^0, y_i;\ \alpha_i, d_i\right), \quad (1)$$

where V_{ij} is the indirect utility for individual i at case j expressed as a function of market prices (p_j), other non-priced attributes (x_j), individual income (y_i), other demographic characteristics (d_i), and latent preferences (α_i).

Ideally, the analyst would estimate willingness to pay using the joint distributions of data describing the observable characteristics of individuals and their choices at policy case *j*, $G_j(y, d)$ and $F_j(p, x, q)$, respectively. If these data are unavailable, however, or if constraints on time and resources prohibit original estimation, then a benefit transfer can be used to estimate wtp_{ij} using preexisting information from a different study case, *k*. We use a *T* superscript to distinguish a benefit transfer estimate for willingness to pay $\left(\widehat{wtp}_{ij}^T\right)$ from what would be obtained from information on policy case *j* if there were no constraints on time, resources, or data $\left(\widehat{wtp}_{ij}\right)$. The difference between these two measures defines the benefit-transfer error (*BTE*):

$$\begin{aligned} BTE\ &= \widehat{wtp}_{ij}^T\left[p_j, x_j, q_j^0, q_j^1, y_i, d_i, F_k(p, x, q), G_k(y, d) | \hat{\beta}_k, v\right] \\ &-\ \widehat{wtp}_{ij}\left[p_j, x_j, q_j^0, q_j^1, y_i, d_i, F_j(p, x, q), G_j(y, d) | \hat{\beta}_j, v\right], \quad (2) \end{aligned}$$

where $\hat{\beta}_k$ is a vector of parameters estimated from the study case data and *v* denotes the valuation methodology (e.g. travel cost, contingent valuation).

The *BTE* is a measure of convergent validity that compares two estimates of the same theoretical value defined in equation (1), which is not an absolute criterion because the true value of wtp_{ij} is unknown. In an ideal setting where the policy case estimator is unbiased, such that $E\left[\widehat{wtp}_{ij}\right] = wtp_{ij}$, the *BTE* will provide an unbiased estimate of $\left(\widehat{wtp}_{ij}^T - wtp_{ij}\right)$. It is important to acknowledge that measurement error and other problems may cause the policy case estimator to be biased, limiting the policy relevance of the *BTE* statistic (Rosenberger and Stanley, 2006). While this issue is important, it is beyond the scope of our study to resolve. Our research design follows the majority of the literature in separating the two problems. That is, we take $\widehat{wtp}_{ij}$ as given and focus on explaining the composite error in measuring $\widehat{wtp}_{ij}$ that is introduced by methodological features of the benefit transfer process and by differences between the study and policy cases.

The general expression for the benefit transfer error in (2) illustrates four distinct ways in which an analyst's modeling choices can influence the magnitude of the *BTE*.[3] First, the size of the *BTE* may depend on how the change in the environmental amenity, $\Delta q = q^1 - q^0$, is defined. Second, errors may stem from systematic differences between the observable characteristics of the study and policy case populations, $G_j \neq G_k$. Third, the *BTE* may vary with the valuation methodology, v. Finally, the error may depend on the transfer procedures embedded in $\widehat{wtp}_{ij}^T$. The individual modeling decisions reflected in the definitions for $\Delta q, v, G,$ and T may interact in ways that increase or decrease the *BTE*. Geographic similarity between the study and policy cases may reduce transfer errors more in value transfers than in function transfers, for example. That is, a function transfer can be calibrated to policy-case conditions by assigning levels to covariates in the transfer equation; whereas no similar calibration is possible for a value transfer so selection of the specific value to transfer that is "similar" becomes critically important.

Transfer errors are typically reported in percentage terms,

$$\text{\% Benefit Transfer Error} = \%BTE = \left[\left(\widehat{wtp}_{ij}^T / \widehat{wtp}_{ij}\right) - 1\right] \times 100. \qquad (3)$$

[3] A fifth potential source of error that is arguably out of the practitioner's control is any systematic difference between the latent preferences of the study and policy case populations.

Studies that do not report $\%BTE$ almost always report sufficient information for readers to make this calculation on their own.

Each study in the convergent validity literature reports one or more transfer errors conditional on a specific set of modeling decisions. We have systematically reviewed these studies and assembled a database documenting transfer errors and transfer characteristics. To extract the signals from the noise we use a meta-regression,

$$|\%BTE| = m(X) + \varepsilon, \tag{4}$$

where X includes a vector of variables that we use to describe analysts' modeling decisions (variables representing $\Delta q, v, G,$ and T).

3. Data Description

3.1. Reviewing the Literature on Convergent Validity of Benefit Transfers

Through an exhaustive search, we identified 40 convergent-validity studies that were published or posted online between 1990 and 2009. Thirty eight of these studies were published in peer-reviewed journals or book chapters. We ultimately excluded the two studies that were not peer reviewed because they had insufficient documentation for some of the key variables ($\Delta q, v, G,$ and T). Three peer reviewed studies were excluded for the same reason. A study by Morrison et al. (2000) was excluded in order to avoid duplication of the findings reported in Morrison et al. (2002). Studies by Engel (2002) and Chattopadhyay (2003) were excluded because benefit transfer errors were reported as ranges rather than point estimates. In addition, a study by Eshet et al. (2007) was excluded because it would have been the only one to use the hedonic methodology, and a single study would not have allowed us to identify the effect of this valuation method on the *%BTE*. These exclusions left us with a total of 31 studies with *%BTE* observations to analyze. Summary statistics describing the transfer applications and results for each included study are reported in Appendix Table I. Citations to all 40 studies and studies excluded from our analysis are provided in Appendix III and Appendix Table II, respectively.

A uniform coding protocol was implemented to ensure that the modeling choices made by the authors of each study, and the corresponding %*BTEs*, were recorded correctly and consistently. The data were double coded by two research assistants, who then met with us to resolve discrepancies. Then we cross checked the coding a second time.

Some studies report what we refer to as “flip” error calculations. The investigators would compute a transfer error with *j* as the policy case and *k* as the study case *a la* equation (3). Then they would flip the two cases, computing a second transfer error with *k* as the policy case and *j* as the study case. Flipping the study and policy cases changes the sign and absolute magnitude of the percentage error in (3). Unfortunately, it was not possible to infer the flip errors for most studies that did not make this calculation directly. Since we were unable to include flip errors for every study, we decided to use a single set of errors from studies that reported flips. Specifically, we used the first set of errors reported by the investigators.[4]

Each of the 31 validity studies reported multiple transfer errors, from as few as 2 to as many as 178. The error magnitudes vary with the transfer characteristics, selection of study cases and study case valuation methods, and amenity of interest. For example, Loomis et al.

[4]Chattopadhyay (2003) proposes an alternative formula for the %*BTE* that would avoid the flip error problem:

$$|\%\ transfer\ error| = \left|\left[\left(\frac{study\ case\ value - policy\ case\ value}{study\ case\ \ vlaue + policy\ case\ value}\right)/2\right] \times 100\right|.$$

This formula produces the same measure of error, regardless of which value is defined as the study case. It would be convenient if future convergent-validity studies were to adopt this metric.

(1995) tested the convergent validity of travel-cost estimates for the average consumer surplus associated with a single day of reservoir-based recreation in Sacramento, CA, Little Rock, AR, and Nashville, TN. After estimating travel-cost models for each of the three regions, they measured benefit transfer errors by making all possible pair-wise comparisons of estimated and transferred consumer surplus for 28 reservoir sites. This yielded 112 distinct transfer errors, half of which were flips. Excluding the flip errors left us with 56 observations on $|\%BTE|$ with a mean of 115%.

The final data set contains 1071 transfer errors. Of these observations, 55% describe applications in Europe, 37% are drawn from United States, and the remaining 8% are from Australia and the rest of the world. [5] The European data include observations from 12 Western European countries and the U.S. data include observations from all of the lower 48 states. The set of applications is also diverse. Eight studies considered the benefits of access to recreation sites (including forest recreation, reservoir based recreation, park recreation, and hunting); five evaluated prospective changes in the quality or quantity of water (including coastal areas, lakes, rivers, and groundwater); and three studies evaluated the benefits of reductions in sources of human health risk (air quality, water quality and ultra violet radiation). Other studies focused on opportunities for fresh water fishing (salmon, trout, big game, small game, flatfish, salmon, steelhead, walleye, pike, bass, and panfish), amenities associated with land preservation (farmland, forested land and coastal land), and the overall ecological health of watersheds, wetlands, and rivers.

[5] We did not find evidence of systematic differences across these four regions in terms of the impact of modeling decisions on benefit transfer errors. Adding fixed effects for regions to the linear meta-regressions did not produce any statistically significant differences in coefficients on benefit transfer characteristics. Unfortunately, there was insufficient variation in the data to identify separate nonparametric models for each region.

3.2. The Distribution of Transfer Errors

Figure 1 illustrates the distribution of absolute benefit transfer errors. In percentage terms, the errors range from 0% to 7,496%, with a mean of 172%. However, the mean reflects a few observations with extremely large errors. The median is 39%, less than a quarter of the mean. The large difference between the mean and the median suggests a need to investigate outlying observations that could influence econometric inferences from the data.

We inspected the data for the presence of outliers using the inter-quartile range (IQR) criterion (Schwertman, Owens and Adnan, 2004). According to this criterion, an observation is classified as an outlier if $|\%BTE| < Q_1 - 1.5 \times IQR$ and/or $|\%BTE| > Q_3 + 1.5 \times IQR$, where $IQR = Q_3 - Q_1$, and Q_1 and Q_3 are the first and third quartiles of the $|\%BTE|$ distribution. This procedure detected 146 outliers.[6] Seventy-two percent of these were reported by just two studies, which also accounted for the largest percentage errors.[7]

Figure 2 graphs the distribution of transfer errors with outliers deleted. The mean and median $|\%BTE|$ are now reduced to 42% and 33%, respectively, and the maximum is reduced to 172% (equal to the mean when outliers were included). One explanation for these large reductions is that some of the validity studies were conducted for situations where the policy case and study case values were conveniently available. These comparisons may be comparisons of conveniently available value data and may not always be considered good candidates for benefit transfers, making it tempting to ignore them. On the other hand, including them in our

[6] We also examined our models for leverage points and influential observations. There were some leverage points, but only a few were found to exert influence on OLS and WLS estimators. However, in general, it is difficult to identify influential and leverage data points when all of the explanatory variables are dichotomous. This observation, combined with the striking spread of the percentage transfer error, is what led us to adopt the IQR criterion.

[7] We considered dropping these two studies, but ultimately decided to keep them in the analysis because most of their observations are not defined as outliers by the IQR criterion (37% are outliers).

analysis may help to identify combinations of procedures that tend to increase benefit transfer errors. To fully evaluate the sensitivity of our results to these large error observations we estimate the model with and without outliers.

3.3. Explanatory Variables

Table 1 defines the variables we use to explain the variation in benefit transfer errors, along with means and standard deviations for each variable. Of the 13 variables in the table, only $|\%BTE|$ is continuous. All 12 explanatory variables are binary indicators describing features of the benefit-transfer applications. They are grouped into the four categories that we defined earlier ($\Delta q, G, v,$ and T). Comparing the means and standard deviations in the last two columns of the table reveals that dropping outliers does not produce any substantial changes in the distribution of explanatory variables.

The first variable in the Δq category, QUALITYΔ, equals one if the transfer describes a change in *quality*, as opposed to a change in *quantity*. For example, changes in human health, river bank erosion, farming practices, air pollution, and water pollution are all defined as *quality* changes, whereas changes in fish catch rates, water supply, and access to recreation sites are all defined as *quantity* changes. The variable USEVALUE indicates whether Δq affects the use value of a resource as opposed to a non-use or total value.

The second category of explanatory variables, *G*, assesses the similarity of the study and policy cases. POPULATION equals 1 if the study and policy case populations are essentially the same. STUDYAREA equals 1 if both cases occur in the same geographic region.

The third set of explanatory variables describes the valuation methodology. Most of the transfer errors are drawn from studies that used choice modeling (CM-31%) or contingent valuation (CV-29%). The remaining observations are based on reduced-form meta-analyses (META-17%), travel cost models of site demand (TC-12%), and random utility models of site choice (RUM-11%).[8]

Finally, we use three variables to describe the nature of the transfer procedures. VALUETRANSFER indicates whether the procedure consisted of a value transfer or a function transfer. MULTIPLESTUDY equals 1 if the study case benefit measure is a composite of results from more than one study case; and MEAN equals 1 if the transfer error is reported as an average over two or more individual transfer errors. [9]

4. Nonparametric Meta-Analysis

4.1. A Nonparametric Model for Binary Regressors

Our nonparametric approach to meta-analysis avoids the need to restrict the functional relationship between benefit-transfer errors and variables describing features of the transfer process. [10] Specifically, we adapt the kernel estimator for models with unordered discrete regressors described in Ouyang, Li and Racine (2009). Our estimation procedure recognizes that

[8] All of the meta-analysis studies use reduced-form regressions, as opposed to the "preference calibration" approach proposed by Smith, Van Houtven, and Pattanayak (2002).

[9] A referee suggested that we consider controlling for variables such as sample size and the number of explanatory variables. The problem with these variables is that they are noisy measures. For example, sample size varies with valuation method (CV observations versus CM choice occasions) and sample frames may be different (e.g., population versus users). Similarly, the range of *%BTE* logically depend on the *relevancy* of the explanatory variables, which may or may not be correlated with the number of explanatory variables. Adding these variables to the model may introduce endogeneity problems that harm, rather than help, the performance of our estimators. Nevertheless, the concept of developing objective measures of the "data quality" of original benefit transfer studies is an interesting topic for future research.

[10] Readers are directed to Boyle, Kaul and Parmeter (2013) for a detailed discussion of the nonparametric methodology used here.

the true form of $m(X)$ in equation (4) is unknown. Equation (4) is estimated using a variant of the Aitchison-Aitken kernel function, which has a simple form and can be easily generalized. To see the mechanics of this process, let λ_r denote a smoothing parameter (or bandwidth) associated with the r^{th} component of *X*. The kernel function for *r* is defined as

$$l(X_{ir}, X_{hr}, \lambda_r) = \begin{cases} 1 & if\ X_{ir} = X_{hr} \\ \lambda_r & if\ X_{ir} \neq X_{hr} \end{cases} , \qquad (5)$$

using $i = 1, 2, \dots, N$ to index individual observations and X_{hr} to denote the value of the r^{th} component of X for the data point that neighbors X_i. The corresponding product kernel can be expressed as

$$L\big(X_{ir,}X_h, \lambda\big) = \prod_{r=1}^{R} l\big(X_{ir}, X_{hr,}\lambda_r\big) = \prod_{r=1}^{R} \lambda_r^{I(X_{ir} \neq X_{hr})}, \qquad (6)$$

where $I(\cdot)$ is an indicator function that equals 1 iff the condition in parentheses is true. Finally, the estimator for the unknown function $m(X)$ is

$$\widehat{m}(X) = \frac{\sum_{i=1}^{N} |\%BTE|_i\, L(X_{ir}, X_h\ \ \lambda)}{\sum_{i=1}^{N} L(X_{ir}, X_h, \lambda)} \ , \qquad (7)$$

The key difference between the nonparametric categorical regression estimator of $m(X)$ and a standard parametric analysis is that the introduction of local weighting allows greater flexibility in uncovering variation in *%BTE*.

To define the effect of a binary variable, let $X_{\sim r}$ denote the subset of regressors in X excluding the r^{th} variable such that $X = [X_{\sim r}, X_r]$. The response of $\widehat{m}(X)$ to changing X_r from 0 to 1 can be written as:

$$\Delta_r = \widehat{m}(0, X_{\sim r}) - \widehat{m}(1, X_{\sim r}), \qquad (8)$$

which can be estimated at every data point. Instead of having a single point estimate for each regressor as in a linear meta-regression, we have a vector of responses that are interpreted as the change in $|\%BTE|$ for a change in X, given the values of the remaining covariates. We refer to these as "response effects".

The rate of convergence for the estimator depends on the relevance of the explanatory variables and on the selection of bandwidths. In the language of nonparametric analysis, an explanatory variable is said to be "irrelevant" if $m(X)$ is constant with respect to that particular variable; i.e. if the response effects are zero everywhere. Hall, Li and Racine (2007) formalize the distinction between relevant and irrelevant variables. To see their distinction, first partition the set of explanatory variables into two components, $X = [\ddot{X}, \dot{X}]$. The variables in $\dot{X}$ are irrelevant, if the dependent variable (Y) and $\ddot{X}$ are independent of $\dot{X}$. The bandwidth for an irrelevant binary variable approaches the upper bound of "1" (Ouyang, Li and Racine, 2009). Therefore, if we find that $\lambda_r \approx 1$, we conclude that *r* is "irrelevant".

We estimate bandwidths using the data-driven least-squares, cross-validation method (LSCV). LSCV employs the leave-one-out technique to find the optimal value of the smoothing parameters. Ouyang, Li and Racine (2009) demonstrate that the rate of convergence to the optimal value of λ is of the order $O_p(n^{-1})$ when all regressors are relevant and $O_p(n^{-0.5})$ when some regressors are irrelevant.[11] Importantly, these convergence rates are faster than nonparametric models with continuous regressors. Thus, the nonparametric model with binary regressors is less vulnerable to the curse of dimensionality. Moreover, if all regressors are

[11] Ouyang, Li and Racine (2009) make an *iid* assumption that is not guaranteed to hold in our data. However, Li, Ouyang and Racine (2013) study the nonparametric kernel regression estimator and optimal bandwidth selection in the presence of weakly dependent data. Their results are similar to the iid results from Ouyang, Li, and Racine. While Li, Ouyang and Racine do not explicitly consider the case where irrelevant variables are included in the regression model, they note, "… we fully expect that the results in Hall et al.'s (2007) analysis carry through to weakly dependent data settings. We present simulation results that support this contention …" (page 699).

discrete and relevant, the nonparametric model converges at the same rate as the correctly specified parametric model.

In the presence of irrelevant variables, the nonparametric model for binary regressors is still consistent, but it converges slower than the correctly specified parametric model. Of course, an econometrician never knows the correct specification for a parametric model. A misspecified parametric model may not converge at all (Ouyang, Li and Racine, 2009).

4.2. Performance in Information Deficient Settings

It is important to consider the space of potential modeling decisions for benefit transfers. With 5 valuation methods and 7 binary regressors describing the transfer process, there are $2^7 \times 5 = 640$ "cells" in the data "grid" describing potential approaches to conducting a benefit transfer.[12] Our data contain 1,071 observations on transfer errors, but these observations are not uniformly distributed across the data grid. Some of the cells are empty. While nonparametric models with binary regressors are not subject to the same "curse of dimensionality" as models with continuous regressors, it may still seem counterintuitive that a nonparametric model would be able to estimate bandwidths with a reasonable degree of accuracy in this environment. Fortunately, there is a wealth of recent evidence that our estimator is, in fact, well suited to "information deficient" environments (Henderson et al., 2012; Ouyang, Li and Racine, 2009; Parmeter, Zheng and McCann, 2009; Savchuk, Hart and Sheather, 2010). We summarize technical details in the appendix.

Of course, the nonparametric model does not allow us to calculate response effects for

[12] The valuation methodologies can be modeled as separate binary variables or as a categorical variable with META as the base category. Both these techniques generate qualitatively similar results. In this paper, we present the results with methodological variables as a category recognizing that we lose the ability to identify separate bandwidths for methodological variables.

empty cells. Greater numbers of empty cells in the support of a particular variable make it harder to determine the full range of impacts of that variable on the benefit transfer error. Adding parametric restrictions to the meta-regression reduces the number of cells that have to be filled in order to identify a variable's impact, while simultaneously increasing the scope for functional form misspecification to bias the estimator. With this tradeoff in mind, we proceed in two stages. First we estimate the meta-regression nonparametrically. The results allow us to characterize ranges of response effects for most variables. Then we estimate a conventional linear meta-regression using weighted least squares, recognizing that the credibility of inferences based on our results is decreasing in the strength of the functional form assumptions we maintain (Manski, 2007).

5. Results

5.1. Nonparametric Model

LSCV bandwidths were estimated separately for the full data set (N=1071) and for the data set excluding outliers (N=925). The two sets of bandwidths differed substantially. Therefore, we adapted Hartarska et al.'s (2010) procedure for dealing with outliers in nonparametric models. Following their logic, we mitigate the influence of outliers by first estimating the optimal bandwidths without outliers and then using the estimated bandwidths to estimate $m(X)$ on the full data set (including outliers). Table 2 reports the resulting LSCV bandwidths, along with summary statistics for the distribution of response effects (mean, median, 25th and 75th quartile).[13] Using the wild bootstrap procedure we estimate cluster robust standard errors to correct for potential heteroskedasticity and correlation of the errors within each original study.[14]

The estimated quartiles of the response effect distributions reveal five clear results. The first result is one of the stylized facts in the literature; value transfers tend to generate larger transfer errors than function transfers. The second result follows logically from what is known about the complexity of research design; transfer errors tend to be larger for studies that consider changes in environmental quality, rather than the quantity, of a particular amenity. Third, contingent-valuation applications generate lower transfer errors than choice modeling. Fourth, travel cost models of site demand are associated with lower transfer errors than random utility models of site choice. Lastly, contingent valuation generates lower transfer errors than meta-analysis, which in turn performs better than choice modeling, travel cost and random utility models. All of these results are large in magnitude, statistically significant, and remarkably robust; the interquartile ranges and means all suggest consistent directions for the response effects. (We will interpret these results in Section 5.3, after summarizing results from the WLS model.)

Other nonparametric response effects are less robust. Convergent validity studies where

[13] We perform a kernel based nonparametric test of significance. This test is analogous to a t-test in linear regression 0. Most of the variables are highly significant except for MEAN that is significant at 10% and USEVALUE, which is insignificant.

[14] Cameron, Gelbach, and Miller (2008) demonstrate that these standard errors perform well in making robust inferences in the presence of clustered observations. After estimating equation (7), a wild bootstrap sample is generated as $\%|BTE|_s^* = \hat{m}(X)_s + \hat{u}_s\epsilon^*, s = 1, \ldots, 31$, where s is the study id and $\epsilon^*$is a white noise term and $\hat{m}(X)$ and $\hat{u}$ are the nonparametric fitted value and residual terms, respectively. The white noise term is defined such that $\mathrm{E}(\epsilon^*) = 0$ and $\mathrm{E}(\epsilon^{*2}) = 1$. Values are randomly selected using a two point distribution given by: $\epsilon^* = \frac{1-\sqrt{5}}{2}$ with probability $p = \frac{1+\sqrt{5}}{2\sqrt{5}}$ and $\epsilon^* = \frac{(1+\sqrt{5}\,)}{2}$ with probability $(1-p)$. The bootstrap sample is then used to estimate new response effects. This procedure is repeated 999 times. Thus, for each estimated response effect of the relevant variables, we will have 999 values. The standard error for an estimated response effect is calculated by taking the standard deviation of the corresponding 999 bootstrapped response effects.

the study and policy cases describe the same geographic area are associated with smaller transfer errors sometimes, but not always. The 25th and 50th quartiles of the response effect distribution are negative, but the 75th quartile is positive. Likewise, in situations where the study case value was defined using data from multiple studies, the negative mean response effect is clearly driven by the observations in the left tail of the distribution of response effects. Finally, the data do not allow us to recover precise estimates for the response effect distributions associated with POPULATION, MEAN, and USEVALUE. The estimated bandwidth for the first variable is 1, and the cluster robust standard errors for the other two variables suggest that their response effect distributions cannot be estimated precisely (MEAN that is significant at 10% and USEVALUE is insignificant). Therefore, in the hope of recovering reasonable approximations to the means of their response effect distributions, we repeat the estimation after adding the conventional linearity and separability restrictions to the meta-regression.

5.2. Parametric Model: Weighted Least Squares

Recall that the 31 convergent validity studies vary considerably in the number of values they report for the %*BTE* (from 2 to 178). Since ordinary least squares estimation assigns equal weight to each observation, studies that provide more observations have greater influence on the results from linear estimation. To mitigate this influence, we estimate equation (4) using weighted least squares (WLS) with cluster robust standard errors. Each observation is weighted by the total number of observations contributed by the corresponding study. Thus, individual observations from studies that provide more observations receive less weight in the estimation. The last two columns of Table 2 report the WLS results, with and without outliers. Not surprisingly, dropping outliers decreases the absolute magnitudes of point estimates for the regression coefficients and improves model fit.[15] The modeling decisions we observe explain three quarters of the variation in the percentage transfer error when outliers are removed.

Imposing the linearity and separability restrictions on the model produces substantial changes in the estimated mean response effects. Nevertheless, with or without outliers, the WLS point estimates are still consistent with most of the qualitative findings from the nonparametric regression. The most striking changes are in the coefficients for the variables describing study case valuation methods. The WLS model suggests that RUM, travel cost, contingent valuation, and choice modeling all tend to produce smaller transfer errors than meta-analysis (the excluded category). While our nonparametric model is consistent with this conclusion for contingent valuation, it provides conflicting evidence that travel cost and meta-analysis produce smaller transfer errors than RUM and choice modeling. A consistent kernel test (Hsiao, Li and Racine, 2007) rejects the null hypothesis that the WLS model is correctly specified against the nonparametric alternative. The differences in sign and magnitude for the mean response effects for RUM, TC, and CM reveal the empirical importance of the misspecification bias in the functional form of the linear model. Moreover, the width of the interquartile range of nonparametric response effects suggests that the linear model overlooks important features of the data.

5.3. Interpretation of the Results

We use a novel approach to graphical analysis—45 degree plots of whisker figures—to help us interpret our findings. The idea for the plots is simple. Both axes of a 2–dimensional diagram are used to represent the same range of response effects for an independent variable

[15] We also estimated the model using OLS. The results were much more sensitive to outliers. This is partly due to the fact that the study with the largest error (7496%) also had the largest number of observations.

(Henderson, Kumbhakar and Parmeter, 2012). This makes it possible to visualize the entire distribution of response effects (and their confidence intervals) along the 45-degree line. To see this, consider Figure 3. The solid square represents our WLS estimate for the impact of value transfers relative to function transfers, using the full data set. Its level (31%) can be seen from the horizontal axis. Its whiskers define a 95% confidence interval around our point estimate, measured on the vertical axis. The solid circle and its whiskers summarize our WLS estimate when we exclude outliers. Each remaining vertical line represents a cluster of nonparametric response effects. The number at the center of each line reports the share (percentage) of effects clustered at that point (e.g. 49% of the effects are clustered at a point near a 200% error). To understand why clustering occurs, recall that all regressors are binary, making $\widehat{m}(X)$ discontinuous. As a result, the response effects tend to be clustered at values that correspond to specific combinations of explanatory variables. When multiple clusters are too dense for their individual shares to be legible, brackets are used to indicate the cumulative share for the group. For example, 26% of response effects are located in a series of clusters just above zero. Finally, the text box indicates that 88% of the response effects are positive.

Visual inspection of the response effect distribution can reveal important features of the data that cannot be seen from the mean response effects and their quartiles in table 2. For example, seeing the right tail of the distribution may be especially important if a benefit transfer practitioner wants to avoid benefit-transfer practices that create the possibility of extreme errors. Seeing the entire distribution also enables the analyst to assess the robustness of WLS mean effects.

5.3.1. Function Transfer versus Value Transfer

Figure 3 confirms the stylized fact that function transfers outperform value transfers (Brouwer and Spaninks, 1999; Rosenberger and Phipps, 2007). Eighty-eight percent of the response effects are in the positive quadrant and so are the point estimates from both WLS models. While moving from WLS to the nonparametric model does not change our conclusion about the sign of the mean response effect, it does suggest two additional findings. First, the superior performance of function transfers is extremely robust. Second, the distribution of response effects has a fat right tail. Moving from function transfer to value transfer often increases the benefit transfer error by as much as 200%. This effect is several times as large as the mean response effects suggested by the WLS models. The superior performance of function transfers is likely due, in part, to the presence of spatial and temporal variation in amenities and variation in household income and preferences (Boyle et al., 2009). Function transfers adjust for features of this heterogeneity, whereas value transfers do not.

5.3.2 Quantity Changes versus Quality Changes

An equally robust result is that benefit transfers are almost always generate lower transfer errors when they describe quantity changes, rather than quality changes. Figure 4 illustrates that 96% of the nonparametric response effects are positive, along with both estimates from the WLS model. Nearly two thirds of the response effects exceed 100%. These findings make sense. Quantity changes are usually easier to describe than quality changes. Anglers can easily understand a change in catch rates or a permanent closure of a fishing site, for example. In contrast, it may be difficult for them to assess a change in water quality, especially when the change is not visible and the effect on fishing is not explicit. Furthermore, consumers may perceive quality differently at the study and policy cases, increasing the transfer error. If the benefit-transfer practitioner is unable to control whether their assessment is framed as a quantity or quality change, the results in Figure 4 suggest that extra caution and additional sensitivity analyses are warranted if the transfer involves a change in environmental quality.

5.3.3 Geographic Similarity

The nonparametric model reveals important features of the response effect distribution for STUDYAREA that are obscured by the linearity and separability assumptions of a conventional meta-analysis. To see this, first notice that our WLS point estimates are both negative, but close to zero (-7% to -14%) and statistically insignificant. In contrast, the mean response effect in the nonparametric model (-52%) is larger and statistically different from zero. It indicates that geographic similarity between the study and policy cases substantially reduces transfer errors, consistent with the stylized facts in the literature (Boyle et al., 2009; Johnston and Rosenberger, 2010; Rosenberger and Phipps, 2007).

However, the mean nonparametric response effect summarizes a wide distribution. While more than half of the response effects are negative, 36% are clustered at a large positive value (nearly 200%) with a very small confidence interval. This counterintuitive cluster of extreme positive values reflects two attributes of the nonparametric analysis. First, the response effects in Figure 5 are only calculated for the observations with STUDYAREA=1.[16] This is a small share of the data (26%). Second, of this 26%, a large cluster of observations come from a single study that happens to have an extreme response effect, which may suggest a poor comparison application was chosen for the validity investigation.

The response effects for STUDYAREA can be divided into two groups – applications with

[16] Similar to dummy variables in a linear model, calculation of response effects requires there to be a base group with a value of 0 for each variable. We defined the base group to be the subset of observations that did not make the modeling decision represented by a '1' for the corresponding variable. If we were to switch the base group, we could calculate response effects for the remaining observations, allowing us to "fill in" the response effect distribution. However, the qualitative results would be unchanged.

value transfers or function transfers. Within each group, the average response effect is negative but the direction of the effect is far more robust for value transfers. Thus, our results reinforce the general importance of geographic similarity between study and policy cases when value transfers are conducted. These results also suggest that function transfers facilitate calibration, based on explanatory variables that might overcome differences in the policy and study cases.

5.3.4 Combining Data from Multiple Study Cases

Intuition would suggest that combining data from multiple study cases would reduce transfer errors as long as the selected study cases are appropriate for the transfer. Moreover, Smith, Van Houtven, and Pattanayak's (2002) logic for "preference calibration" suggests there are gains from using multiple study cases to span the relevant portion of the policy case benefit function. Indeed, the WLS models suggest a modest reduction in *%BTE* and the mean nonparametric response effect for MULTIPLESTUDY in Table 2 appears to confirm this finding.

Figure 6 reveals some interesting heterogeneity underlying the mean response effect. Just over three quarters of the effects are positive, but they are all less than 1% with tight confidence intervals. The negative response effects are at least an order of magnitude larger; the mean is -28% and the maximum is -9%. Focusing on the negative response effects, 86% correspond to function transfers and 73% correspond to studies that evaluated quantity changes. In contrast, almost all of the value transfers have positive (but near zero) response effects. These findings indicate that combining values from several studies has greater scope to improve the accuracy of function transfers that can adjust for differences between the sites.

5.3.5. Valuation Methodology

The nonparametric response effects for study case value methods are the most difficult to interpret. Designs with more choice alternatives (e.g., choices in CM and sites in RUM) appear to be associated with larger errors. In the context of stated preference research, contingent valuation is associated with smaller errors than choice modeling. On the revealed preference side, travel cost models of site demand tend to have smaller errors than random utility models of site choice. The mean response effects range from -70% for CV to 349% for CM. Meta-analysis lies in between these extremes, which is not surprising since the meta-analyses in our data are mostly based on study case value estimates derived from two or more of these four valuation methods.

One interpretation of these results is that increasing the number of choice alternatives leads to more complex econometric models, increasing the scope for modeling decisions to introduce bias and increase the transfer error.[17] However, we must be extremely cautious with this interpretation. It is important to keep in mind that, on average, CV studies generate the fewest transfer errors per study.[18] In comparison, CM and META studies often test convergent validity by comparing each transfer estimate with each of the remaining 'N-1' alternatives. The "extra" transfer errors generated by CM, RUM, META and TC based studies may be econometric targets of opportunity that would not be considered best practices in a real world policy evaluation.

More importantly, our measure for the benefit transfer error does not say anything about

[17] As Boyle et al. (2009) demonstrate, specification of a utility function for an econometrically consistent RUM or CM transfer requires that four "S" assumptions are satisfied. The study site model must be correctly *specified*; unobserved attributes of the study and policy sites must be *separable* from observed attributes in the utility function; consumers must not be *sorted* across the study and policy sites according to unobserved attributes of their preferences; and the data on demographic attributes of policy site consumers must not present any form of *selection* bias.

[18] A total of twelve contingent valuation studies generate 29% of the benefit transfer errors, therefore, the average transfer errors per study is about 2.4%. The average transfer errors per study exceed 3% for the rest of the study case valuation methods.

the relative performance of these methods in estimating the actual wtp at the policy and study cases using case-specific data. At most our results only illustrate that CV has performed relatively better in benefit transfer applications than other methods. Following this result, an important area for future research would be to compare the performance of study case valuation methodologies in the context of benefit transfer. This can be achieved by conducting a benefit transfer exercise for the same set of study and the policy cases using more than one valuation method.

6. Conclusions

What are the practical implications from the past 20 years of research on the convergent validity of benefit transfers? Out of 1071 transfer errors reported by 31 studies, the median absolute error is 39% for all data and 33% with the outliers removed. Individual transfer errors range from close to 0% to larger than 200%. Both the median transfer error and its range are similar in magnitude to certain types of errors associated with original studies. For example, Murphy et al. (2005) assess the importance of "hypothetical bias" in stated preference research by analyzing results from 28 laboratory and field experiments on the difference between hypothetical statements of value and actual statements of value based on binding commitments. The median hypothetical value overstated actual value by 35%, with the majority of estimates for hypothetical bias falling between 0% and 200%.

Our results demonstrate characteristics of benefit transfers that are associated with greater or lesser errors. These insights can help to guide future analyses on where caution is important in the conduct of a benefit transfer and robustness checks are warranted. The evidence overwhelmingly supports the stylized fact that function transfers outperform value transfers. The literature also suggests that benefit transfers are better able to predict the willingness to pay for quantity changes than for changes in environmental quality. The former is a choice made by the benefit-transfer practitioner, while the latter is defined by the policy question being addressed. That said, if the practitioner must perform a value transfer then it becomes especially important to ensure that the study-case value estimate matches the policy-case value definition. We also find that using information from multiple and geographically similar study cases can reduce benefit transfer errors. Site similarity tends to be more important for value transfers, but function transfers appear better equipped to exploit information from multiple studies.

It is important to distinguish between the way that meta-analysis is used for benefit transfers and the way that we have used the methodology in this study. We have developed a new nonparametric approach to meta-analysis and demonstrated that it can extract important signals from the data that remain hidden in conventional linear models. This advancement continues a long tradition of refining the econometrics of meta-analysis in order to distill key findings from research on important questions in environmental economics (Banzaf and Smith, 2007; Smith and Kaoru, 1990; Smith and Osborne, 1996; Walsh, Johnson and McKean, 1989). In contrast, when meta-analysis is used to conduct a benefit transfer the methodology must provide policy-relevant transfer estimates. Parametric estimation is desirable for providing point estimates of meta-equations parameters, but the robustness of these parameter estimates should be investigated using nonparametric estimation. Our results on the performance of methodological variables need to be carefully interpreted. For benefit transfer, we find that contingent valuation and travel cost models perform better than choice modeling and random utility models, respectively. However, these results do not suggest anything about the relative performance of these methods in estimating the actual wtp at the policy and study cases using case-specific data. An important area for future investigation would be to test the performance of benefit transfer applications for specific valuation methods (e.g., CV, RUM, META, etc.) holding the study and policy case applications constant. Our results on META indicate that the performance of meta-analysis based on specific valuation methods (e.g. CV and TC), in the context of benefit transfer, should be investigated in the future.

Finally, we have some recommendations for future studies evaluating the validity of benefit transfers:

- In our review of the literature, we observed several cases where transfer characteristics were not clearly documented. Future benefit-transfer validity analyses must be more thorough in their documentation of data and analysis procedures.
- Wide ranges of transfer errors were sometimes presented, but not explained. A documentation protocol is needed in order for future convergent-validity studies to enhance the credibility of benefit transfers. Each study should define the criteria they use to identify study and policy cases that are good candidates for benefit transfers, and then justify each comparison in the context of those criteria.
- A protocol for computing errors should be used, such as presented by Chattopadhyay (2003), so that benefit transfer errors are invariant to the specification of study and policy cases.
- Investigators need to go beyond simply reporting transfer errors to explore why some comparisons have small errors and others have large errors.

Taking these steps would provide insights to refine the criteria for when appropriate data are available to conduct benefit transfers and enhance the credibility of benefit transfers.

References

H.S. Banzaf, V.K. Smith. 2007. Meta-Analysis in Model Implementation: Choice Sets and the Valuation of Air Quality Improvements. Journal of Applied Econometrics, 22 (6): 1013-1031.

J.C. Bergstrom, L.O. Taylor. 2006. Using Meta-Analysis for Benefits Transfer: Theory and Practice. Ecological Economics, 60(2): 351-360.

K.J. Boyle, N.V. Kuminoff, C.F. Parmeter, J.C. Pope. 2009. Necessary Conditions for Valid Benefit Transfers. American Journal of Agricultural Economics, 91: 1328-1334.

K.J. Boyle, N.V. Kuminoff, C.F. Parmeter, J.C. Pope. 2010. The Benefit-Transfer Challenges. Annual Review of Resource Economics, 2 (1): 161-182.

K. J. Boyle, S. Kaul, C. F. Parmeter. 2013. Meta-Analysis – Advances And New Perspectives Toward Data Synthesis And Estimation Robustness, in R. Johnston, J. Rolfe, R. Rosenberger, R. Brouwer (eds.), Benefit Transfer of Environmental and Resource Values: A Handbook for Researchers and Practitioners, forthcoming.

Brouwer, R., F. A. Spaninks. 1999. The Validity of Environmental Benefits Transfer: Further Empirical Testing. Environmental and Resource Economics, 14 (1): (95-11).

Cameron, A. C., J. B. Gelbach, D. L. Miller. 2008. Bootstrap-Based Improvements for Inference with Clustered Errors. The Review of Economics and Statistics, 90(3): 414–427.

Chattopadhyay, S. 2003. A Repeated Sampling Technique in Assessing the Validity of Benefit Transfer in Valuing Non-Market Goods. Land Economics, 79 (4): 576-596.

Engel, S. 2002. Benefit Function Transfers Versus Meta-Analysis as Policy Making Tools: A

Comparison, in: R. J. G. M. Florax, Peter Nijkamp, Kenneth George Willis (eds.), Comparative Environmental Economic Assessment, Edward Elgar, Cheltenham, UK.

Eshet, T., M.G. Baron, M. Shechter. 2007. Exploring Benefit Transfer: Disamenities of Waste Transfer Stations. Environmental and Resources Economics 37(3): 521-547.

Hall, P., Q. Li, J. Racine. 2007. Nonparametric Estimation of Regression Functions in the Presence of Irrelevant Regressors. The Review of Economics and Statistics, 89(4): 784–789.

Hartarska, V., C. F. Parmeter, D. Nadolnyak, B. Zhu. 2010. Economies of Scope for Microfinance: Differences across Output Measures. Pacific Economic Review, 15 (4): 464–481.

Henderson, D. J., S. C. Kumbhakar, C. F. Parmeter. 2012. A Simple Method to Visualize Results in Nonlinear Regression Models. Economic Letters, 117 (3): 578 – 581.

Henderson D. J., C. Papageorgiou, C. F. Parmeter. 2012. Growth Empirics without Parameters. The Economic Journal, 122 (559): 125-154.

Hsiao, C., Q. Li, J. Racine. 2007. A Consistent Model Specification Test with Mixed Discrete and Continuous Data. Journal of Econometrics, 140(2): 802–826.

Johnston, R.J. 2007. Choice Experiments, Site Similarity and Benefits Transfer. Environmental and Resources Economics, 38(3): 331-351.

Johnston, R.J., R.S. Rosenberger. 2010. Methods, Trends and Controversies in Contemporary Benefit Transfer. Journal of Economic Surveys, 24 (3): 479-510.

Kirchhoff, S., B.G. Colby, J.T. LaFrance. 1997. Evaluating the Performance of Benefit Transfer: An Empirical Inquiry. Journal of Environmental Economics and Management, 33 (1): 75-93.

Li, Q., D. Ouyang, J. S. Racine. 2013. Categorical Semi Parametric Varying-Coefficient Models. Journal of Applied Econometrics, 28 (4): 551 - 579.

Loomis, J.B., B. Roach, F. Ward, R. Ready. 1995. Testing Transferability of Recreation Demand Models Across Regions: A Study of Corps of Engineer Reservoirs. Water Resources Research, 31 (3): 721–730.

Loomis, J.B., R. S. Rosenberger. 2006. Reducing Barriers in Future Benefit Transfers: Needed Improvements in Primary Study Design And Reporting. Ecological Economics, 60: 372-378.

Manski, C.F. 2007. Identification for Prediction and Decision. Harvard University Press, Cambridge.

Moeltner, K, R. S. Rosenberger. 2008. Meta-Regression and Benefit Transfer: Data Space, Model Space, and the Quest for 'Optimal Scope'. B. E. Journal of Economic Analysis and Policy, 8(1): Article 31.

Morrison, M., J. Bennett, R. Blamey, J. Louivere. 2000. Choice Modeling, Non-Use Values and Benefit Transfer. Economic Analysis and Policy, 30 (1): 13-32.

Morrison, M., J. Bennett, R. Blamey, J. Louviere. 2002. Choice Modeling and Tests Of Benefit Transfer. American Journal of Agricultural Economics, 84 (1): 161-170.

Murphy, J. J., P. G. Allen, T. H. Stevens, D. Weatherhead. 2005. A Meta-Analysis of

Hypothetical Bias in Stated Preference Valuation. Environmental and Resource Economics, 30(3): 313–325.

Nelson, J.P., P.E. Kennedy. 2009. The Use (and Abuse) of Meta-Analysis in Environmental And Natural Resource Economics: An Assessment. Environmental and Resource Economics 42(3): 345-377.

Ouyang, D., Qi Li, J.S. Racine. 2009. Nonparametric Estimation of Regression Functions with Discrete Regressors. Econometric Theory 25(1): 1-42.

Parmeter, C. F., Z. Zheng, P. McCann. 2009. Cross-Validated Bandwidths and Significance Testing, in Q. Li and J. S. Racine (eds.), Advances in Econometrics: Nonparametric Methods, Elsevier Science.

Racine, J. S., J. Hart, Q. Li. 2006. Testing the Significance of Categorical Predictor Variables in Nonparametric Regression Models. Econometric Reviews 25: 523-544.

Rosenberger, R. S., T. Phipps. 2007. Correspondence and Convergence in Benefit Transfer Accuracy: Meta-Analytic Review of The Literature, in S. Navrud and R. Ready (eds.), Environmental Value Transfer: Issues and Methods.

Rosenberger, R. S., T. D. Stanley. 2006. Measurement, Generalization, and Publication: Sources of Error in Benefit Transfers and Their Management. Ecological Economics 60 (2): 372–378.

Savchuk, O. Y., J. D. Hart, S. J. Sheather. 2010. Indirect Cross-Validation for Density Estimation. Journal of the American Statistical Association, 105 (489): 415-424.

Schwertman, N. C., M. A. Owens, R. Adnan. 2004. A Simple More General Boxplot Method for

Identifying Outliers. Computational Statistics & Data Analysis 47 (1): 165–174.

Smith, V.K., G. Van Houtven, S. K. Pattanayak. 2002. Benefit Transfer via Preference Calibration: 'Prudential Algebra' for Policy. Land Economics 78 (1): 132-152.

Smith, V.K., S. K. Pattanayak. 2002. Is Meta-Analysis a Noah's Ark for Non-Market Valuation. Environmental and Resource Economics, 22: 271-296.

Smith, V.K., Y. Kaoru. 1990. Signals or Noise? Explaining the Variation in Recreation Benefit Estimates. American Journal of Agricultural Economics, 72(2): 419-433.

Smith, V.K., L. Osborne. 1996. Do Contingent Valuation Estimates Pass a Scope Test? A Meta Analysis. Journal of Environmental Economics and Management, 31: 287-301.

U.S. Environmental Protection Agency. 2010. Guidelines for Preparing Economic Analyses. EPA 240-R_10-001 (prepublication edition).

http://yosemite.epa.gov/ee/epa/eed.nsf/pages/Guidelines.html/$file/Guidelines.pdf, accessed December 31, 2010.

U.S. Executive Office of the President. 1993. Executive Order 12866: Regulatory Planning and Review. Federal Register 58(190): 51735- 51744.

Walsh, R. G., D. M. Johnson, J. R. McKean. 1989. Issues in Nonmarket Valuation and Policy Application: A Retrospective Glance. Western Journal of Agricultural Economics, 14(1): 178-188.

Table 1: Variables, Definitions and Summary Statistics

Variables	**Definitions**	**All Data**	**Without Outliers**
		Mean **(Standard Deviation)**	
|%Benefit Transfer Error|	$\left\|\left[\left(\frac{\widehat{wtp}_{ij}^{T}}{\widehat{wtp}_{ij}}\right)-1\right]\times 100\right\|$	171.97 (555.19)	42.39 (36.39)
Amenity (Δq)			
QUALITYΔ	1 if a change in quality; 0 if change in quantity	0.47 (0.50)	0.45 (0.50)
USEVALUE	1 if use value; 0 if non-use value	0.66 (0.47)	0.69 (0.46)
Study and Policy Cases (G)			
POPULATION	1 if study and policy case populations are the same; 0 otherwise	0.09 (0.29)	0.11 (0.31)
STUDYAREA	1 if study and policy cases geographic area are the same; 0 otherwise	0.26 (0.44)	0.24 (0.43)
Valuation Methodology (v) *(META is the base/omitted category in estimation)*			
META	1 if the valuation method is a meta-analysis	0.17 (0.38)	0.19 (0.39)
RUM	1 if the valuation method is a random-utility model	0.11 (0.32)	0.09 (0.29)
TC	1 if the valuation method is a travel-cost model	0.12 (0.32)	0.12 (0.32)
CV	1 if the valuation method is contingent valuation	0.29 (0.45)	0.33 (0.47)
CM	1 if the valuation method is choice modeling	0.31 (0.46)	0.27 (0.44)
Transfer Procedures (T)			
VALUETRANSFER	1 if value transfer; 0 if function transfer	0.38 (0.48)	0.36 (0.48)
MULTIPLESTUDY	1 if two or more study cases are used to estimate study-case value; 0 otherwise	0.27 (0.45)	0.31 (0.46)
MEAN	1 if transfer error is reported as a mean of two or more transfer errors; 0 otherwise	0.15 (0.36)	0.17 (0.37)
N	Number of observations	1071	925

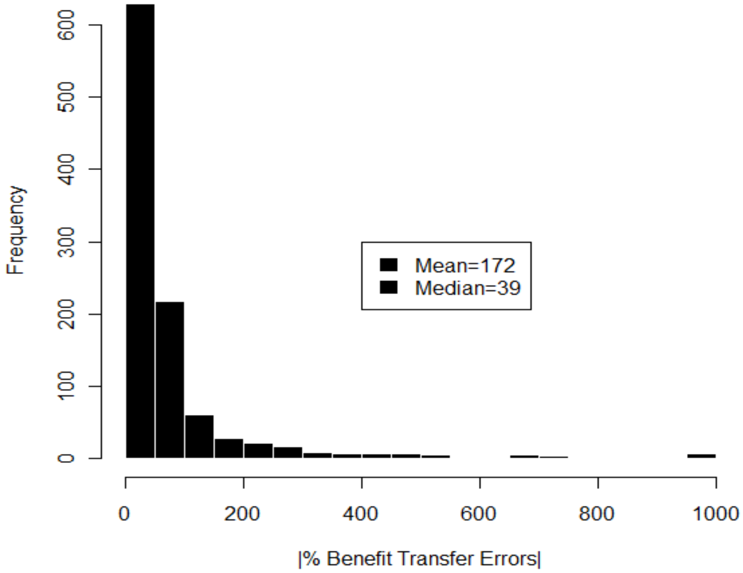
Table 2: Nonparametric and Parametric Meta-Regression Results

		Nonparametric Regression[a]				WLS[b]	
	BW	**Response Effect Quartiles**				**All Data**	**Without Outliers**
		Mean	**25%**	**50%**	**75%**		
QUALITYΔ	0.00	263.36 (135.50)	27.04 (0.54)	172.72 (57.73)	590.85 (69.32)	49.84* (26.27)	24.85* (13.09)
USEVALUE	0.15	-3.43 (1.48)	-8.44 (6.47)	-8.44 (6.47)	-2.13 (7.68)	26.33 (18.28)	4.25 (9.76)
POPULATION	1	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	37.79 (38.56)	19.82 (18.20)
STUDYAREA	0.00	-51.86 (17.09)	-247.29 (115.43)	-70.33 (38.72)	194.22 (0.19)	-14.17 (15.22)	-7.10 (10.85)
RUM	0.00	172.36 (0.20)	194.23 (0.20)	194.23 (0.20)	194.23 (0.20)	-132.97*** (48.75)	-60.34*** (23.28)
TC	0.00	18.82 (0.19)	6.13 (0.46)	27.00 (43.07)	27.00 (43.07)	-130.62** (51.21)	-68.48*** (24.60)
CV	0.00	-70.17 (17.08)	-156.34 (2 6.26)	-37.70 (18.10)	-37.70 (18.10)	-150.10*** (36.95)	-65.94*** (16.56)
CM	0.00	349.71 (135.50)	172.72 (57.73)	329.62 (135.50)	503.10 (69.97)	-91.60** (36.30)	-29.89* (16.28)
VALUETRANSFER	0.41	121.79 (91.37)	1.17 (1.96)	156.91 (91.37)	202.61 (87.16)	31.22* (17.44)	4.79 (4.96)
MULTIPLESTUDY	0.02	-8.80 (0.64)	0.19 (2.81)	0.19 (2.81)	0.19 (2.81)	-23.00** (9.76)	-9.66 (6.62)
MEAN	0.00	-24.61 (17.09)	-15.14 (18.10)	-15.14 (18.10)	-15.14 (18.10)	-21.44 (15.23)	-9.24 (10.67)
Intercept						126.42*** (38.88)	77.02*** (18.16)
$\bar{R}^2$						0.33	0.76
F - value						47.95	269.72
N	925	1071	1071	1071	1071	1071	925

[a] The variable META is excluded from this table as it is the base category for valuation methods. Bandwidths are estimated with outliers excluded. Robust cluster wild bootstrapped standard errors are given in parentheses. Bandwidths for RUM, CV, TC, and CM are identical as they included as one categorical variable with META as the base category. The fact that the 25th, 50th and 75th quartiles in the distribution of response effects are identical for some explanatory variables reflects the discreteness in variable space. The bandwidths for variables QUALITYΔ, STUDYAREA, RUM, CV, TC, CM, and MEAN suggest that their response effects are generated from a frequency type estimation. The intercept for nonparametric model can be easily computed, however, they are not useful for our analysis and are not reported them in this table. [b] Significance codes for WLS models: '***' 0.01, '**' 0.05, '*' 0.1.

Figure 1: Distribution of Percentage Transfer Errors[a]

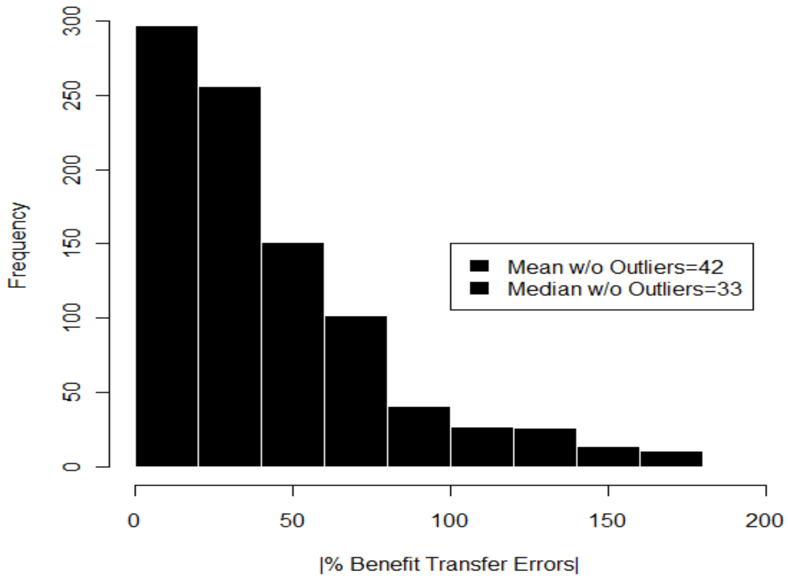
(N=1071)



[a] The scaling of the horizontal axis excludes 3.83% of observations with errors exceeding 1000%.

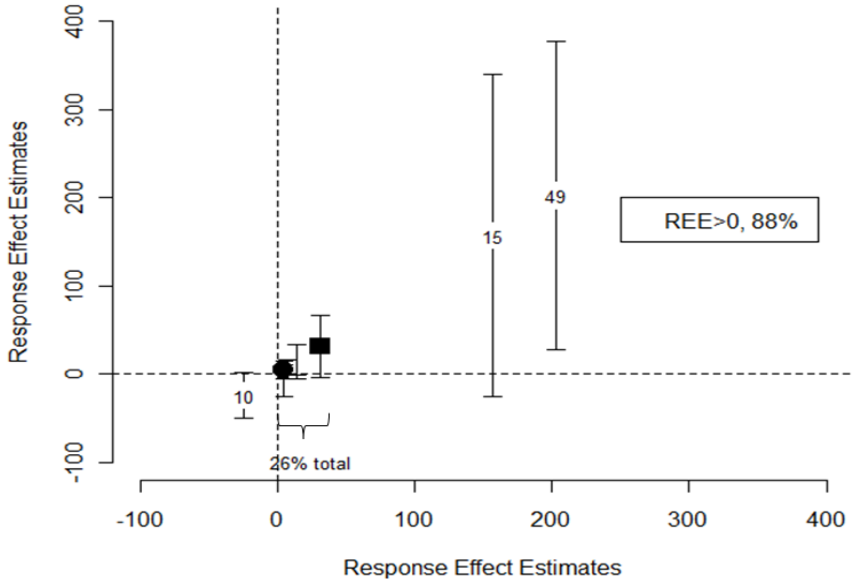
Figure 2: Truncated Distribution of Percentage Transfer Errors, Excluding Outliers [a]

(N= 925)



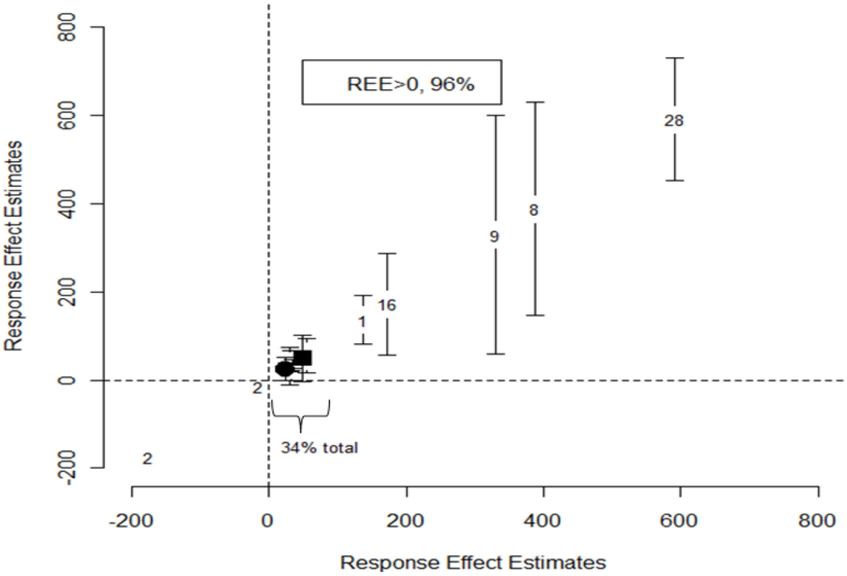
[a] This histogram is drawn for $|\%BTE|$ without outliers. Of 1071 observations 13.63% are outliers according to the interquartile range criterion.

Figure 3: Benefit Transfer Error Response Effects for VALUETRANSFER[a]



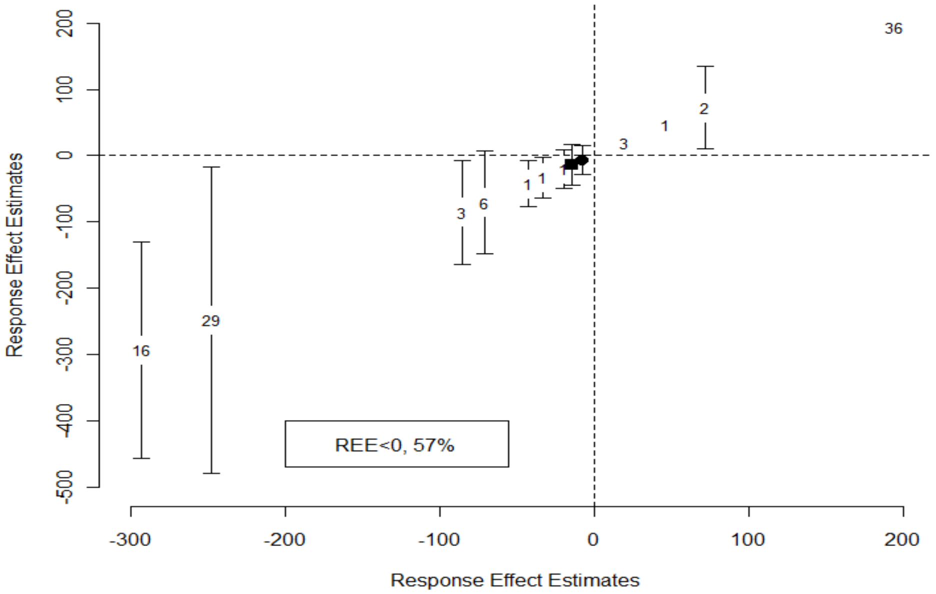
[a] The figure plots WLS point estimates for the VALUETRANSFER variable, nonparametric response effect estimates (REE), and 95% confidence intervals for each. The square and circle represent the WLS point estimates based on the data with and without outliers, respectively. The horizontal bars above and below denote 95% confidence intervals. The numbers indicate the share of the response effects at the point where the number is located. Clustering of REEs occurs because all of the independent variables are binary.

Figure 4: Benefit Transfer Error Response Effects for QUALITYΔ[a]



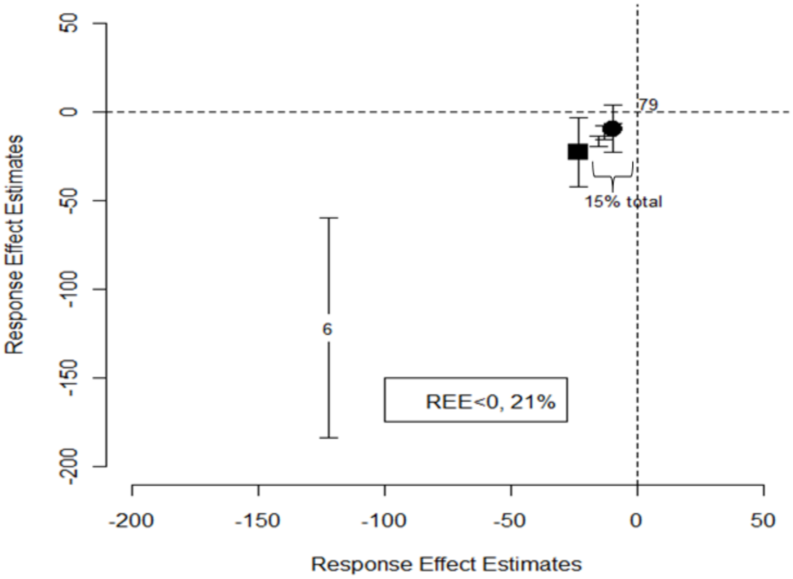
[a] The figure plots WLS point estimates for the QUALITYΔ variable, nonparametric response effect estimates (REE), and 95% confidence intervals for each. Response effects in the negative quadrant have very small standard errors resulting into small confidence intervals, which are not legible on this graph. See the footnote to Figure 3 for additional explanation.

Figure 5: Benefit Transfer Error Response Effects for STUDYAREA[a]



[a] The figure plots nonparametric response effect estimates (REE), and 95% confidence intervals for the STUDYAREA variable. Majority of the response effects in the positive quadrant have very small standard errors resulting into small confidence intervals that are not legible on this graph. See the footnote to Figure 3 for additional explanation.

Figure 6: Benefit Transfer Error Response Effects for MULTIPLESTUDY[a]



[a] The figure plots nonparametric response effect estimates (REE), and 95% confidence intervals for the MULTIPLESTUDY variable. 79% of response effects in the positive quadrant are valued at 0.19%. These response effects have very small standard errors resulting into small confidence intervals that are not legible on this graph. See the footnote to Figure 3 for additional explanation.

Appendix 1. Nonparametric Estimation in Information Deficient Settings

A well-known paradox in the theoretical study of nonparametric cross validation methods is that "the harder the estimation problem, the better cross-validation performs" (see Savchuk, Hart and Sheather (2010) for a summary and citations to the literature). This paradox is exemplified by the 'information deficient' setting in which we are working. In Theorems 2.1 and 3.1 of their paper, Ouyang, Li, and Racine (2009) demonstrate that the convergence rate of the bandwidth for our estimator does not depend on the number of empty cells. Thus, the presence of empty cells will not deteriorate the empirical performance of cross-validation relative to the theoretical insights.

Recent studies have also produced simulation-based evidence on the performance of both the nonparametric estimator and the cross-validated bandwidths. While none of the simulation environments provide an exact match to our data, some have considered problems that are *harder* than the one we face. The strongest evidence can be found in Henderson, Papageorgiou and Parmeter (2012). Their simulations investigate model performance in a setting with 10 covariates (5 continuous, 5 discrete) using sample sizes ranging from 409 to 719. They find that cross validation performs well (in terms of the bandwidths) and that the nonparametric estimator outperforms misspecified linear models. Importantly, the mixed discrete-continuous nature of their data would imply slower rates of convergence for the bandwidths than in our study. Parmeter, Zheng and McCann (2009) report similar results in terms of the cross-validated bandwidths' ability to correctly smooth out irrelevant variables. Thus, the collective evidence from the recent literature suggests that having a large number of covariates and empty cells does not prevent the cross-validated bandwidths from being able to distinguish relevant from irrelevant covariates; nor does it prevent the estimator from identifying the response effects for explanatory variables in the cells that are filled.

Appropriate selection of the smoothing parameter also alleviates the information deficiency issue by combining empty or sparsely filled cells with nearby cells containing data thereby leveraging information from cells that are deemed “close”. Consider the category of study case valuation method, “v”. We combine the variables in this category (META, RUM, TC, CV and CM) into a single categorical variable taking values in (0, 1, 2, 3, 4). In the presence of many discrete variables combining variables that define the same procedure is a viable strategy for improving the success of bandwidth estimation; however, we dispense the ability to identify separate bandwidths for each of these variables as they are combined into a single category. Although we cannot identify separate bandwidths, Ouyang, Li and Racine (2009) demonstrate this will yield qualitatively similar insights into each cell’s “effect” on the response of interest implying that we can consistently estimate each cell’s effect.

Now suppose that no observations exist for category ‘2’. In this case we can still present function estimates using the smoothing approach. Our nonparametric categorical regression estimator for category ‘2’ would be:

$$\widehat{m}(2) = \frac{\sum_{i=1}^{n} |\%BTE|_i L(X_i,2,\lambda)}{\sum_{i=1}^{n} L(X_i,2,\lambda)} = \frac{\sum_{|X_i-2|=1} |\%BTE|_i\lambda + \sum_{|X_i-2|=2} |\%BTE|_i\lambda^2}{\sum_{|X_i-2|=1}\lambda + \sum_{|X_i-2|=2}\lambda^2} \qquad (9)$$

where λ is the bandwidth used to smooth over the five cells. This shows that the nonparametric smoothing estimator is a weighted average of nearby cells (0, 1, 3, 4). The weight depends upon the distance of a cell from the cell of interest, $|X_i - 2|$. In the parametric approach, there would only be four categories (0, 1, 3, 4), and the category ‘2’ would have been placed in the baseline category ‘0’, meaning that the function estimate for true ‘0’ members would also be applied to

‘2’ members. However, in case of nonparametric regression, the estimator leverages information from the nearby cells.

Appendix 2. Characteristics of 31 Benefit Transfer Validity Studies Included in the Meta Analysis

Authors (year)	Valuation Methods	Resources Valued	N	% transfer error	|% transfer error|
				mean (min, max)	mean (min, max)
Barton (2002)	CV	Beach Quality	8	-20 (-23,-10)	20 (10,23)
Bergland, Magnussen and Navrud (1995)	CV	Water Quality	2	-21 (-24,-18)	21 (18,24)
Brouwer and Bateman (2005)	CV	Human Health	85	17 (-41,123)	34 (0.4,123)
Brouwer and Spaninks (1999)	CV	Farm Land	8	3 (-59,60)	42 (22,60)
Colombo and Hanley (2008)	CM	Farm Land	178	680 (2, 7496)	680 (2,7496)
Colombo, Calatrava-Requens, and Hanley (2007)	CM	Soil Conservation	54	110 (8,1148)	110 (8,1148)
Groothuis (2005)	CV/TC	Deer Hunting	120	-10 (-75,136)	30 (0.1,136)
Hanley, Wright, and Alvarez-Farizo (2006)	CM	Ecosystem Health	2	-72 (-78,-67)	72 (67,78)
Jiang, Swallow, and McGonagle (2005)	CM	Coastal Land	5	-68 (-85,-53)	68 (53,85)
Johnston and Duke (2009)	CM	Farm Land	4	-76 (-100,-29)	76 (29,100)
Johnston (2007)	CM	Mixed Resources	24	-12 (-101,58)	37 (7,101)
Kerr and Sharp (2006)	CM	Ecosystem Health	22	79 (-63,704)	120 (2,704)
Kristofersson and Navrud (2007)	CV	Fishing/ Ecosystem Health	21	125 (7,319)	125 (7,319)
Lindhjem and Navrud (2008)	META	Multiple Use Forestry	16	73 (25,266)	73 (25,266)
Loomis et al. (1995)	TC	Reservoir	56	106	115

				(-50,475)	(0.5,475)
Loomis (1992)	TC	Sport Fishing	10	0.2 (-18,9)	6 (1, 18)
Matthews, Hutchinson, and Scarpa (2009)	CV	Forest Recreation	84	12 (-42,125)	27 (0.0,125)
Morrison and Bennett (2006)	CM	Ecosystem Health	28	-25 (-171,30)	35 (1,171)
Morrison et al. (2002)	CM	Wetlands	9	-32 (-66,-4)	32 (4,66)
Parsons and Kealy (1994)	RUM	Water Recreation	11	-4 (-66,75)	21 (1,75)
Piper and Martin (2001)	CV	Water Supply	8	35 (-9,149)	39 (3,149)
Ready and Navrud (2007)	CV	Human Health	2	39.1 (37.7,41.9)	39.1 (37.7,41.9)
Ready et al. (2004)	CV	Human Health	21	37 (20,83)	37 (20,83)
Rosenberger and Loomis (2000)	META	Mixed Resources	115	17 (-79,319)	49 (0.0,319)
Rozan (2004)	CV	Air Quality	4	-27.5 (-28,-27)	27.5 (27,28)
Shrestha and Loomis (2003)	META	Mixed Resources	34	60 (-74,411)	84 (12,411)
Shrestha and Loomis (2001)	META	Outdoor Recreation	18	6 (-46,81)	28 (0.5,81)
Stapler and Johnston (2009)	META	Sport Fishing	4	228 (64,572)	228 (64,572)
Vandenberg, Poe and Powell (2001)	CV	Ground Water	8	29 (16,44)	29 (16,44)
Zanderson, Termansen and Jensen (2007a)	RUM	Forest Recreation	6	4 (-66,55)	30 (4,66)
Zanderson, Termansen and Jensen (2007b)	RUM	Forest Recreation	104	180 (-73,1111)	194 (1.3,1111)

Appendix 3. Validity Studies Excluded from the Meta-Analysis

Authors (year)	Reason For Excluding
Bowker, English and Bergstrom (1997)	Insufficient documentation of key variables
Chattopadhyay (2003)	No point estimate for transfer error
Downing and Ozuna (1996)	Insufficient documentation of key variables
Engel (2002)	No point estimate for transfer error
Eshet, Baron and Shechter (2007)	Only study to use hedonic methodology
Jeong and Haab (2004)	Insufficient documentation of key variables
Kirchhoff, Colby and LaFrance (1997)	Insufficient documentation of key variables
Leon et al. (2002)	Insufficient documentation of key variables
Morrison et al. (2000)	Redundant, given Morrison et al. (2002)

Appendix 4. Reference List for Benefit Transfer Convergent Validity Studies

Barton, D.N. 2002. The Transferability of Benefit Transfer: Contingent Valuation of Water Quality Improvements in Costa Rica. Ecological Economics, 42 (1-2) (2002) 147-164.

Bergland, O., K. Magnussen, S. Navrud. 2002. Benefit Transfer: Testing for Accuracy and Reliability, in: R.J.G.M. Florax, P. Nijkamp, K.G. Willis (Eds.), Comparative Environmental Economic Assessment, Edward Elgar, Cheltenham, UK.

Bowker, J.M., D.B.K. English, J.C. Bergstrom. 1997. Benefits Transfer and Count Data Travel Cost Models: An Application and Test of a Varying Parameter Approach with Guided Whitewater Rafting. Faculty Series (16703), Department of Agricultural and Applied Economics, University of Georgia, Athens.

Brouwer, R., F.A. Spaninks. 1999. The Validity of Environmental Benefits Transfer: Further Empirical Testing. Environmental and Resource Economics, 14 (1): 95-117.

Brouwer, R., I. J. Bateman. 2005. Benefits Transfer of Willingness to Pay Estimates and Functions for Health-Risk Reductions: A Cross-Country Study. Journal of Health Economics, 24 (3): 591–611.

Chattopadhyay. S. 2003. A Repeated Sampling Technique in Assessing the Validity of Benefit Transfer in Valuing Non-Market Goods. Land Economics, 79 (4): 576-596.

Colombo, S., N. Hanley. 2008. How Can We Reduce the Errors From Benefits Transfer? An Investigation Using the Choice Experiment Method. Land Economics ,84 (1): 128-147.

Colombo, S., J. Calatrava-Requens, N. Hanley. 2007. Testing Choice Experiment for Benefit Transfer with Preference Heterogeneity. American Journal of Agricultural Economics, 89 (1): 135-151.

Downing, M., T. Ozuna. 1996. Testing the Reliability of The Benefit Function Transfer Approach. Journal of Environmental Economics and Management, 30 (3): (1996) 316-322.

Engel, S. 2002. Benefit Function Transfers Versus Meta-Analysis as Policy Making Tools: A Comparison, in: R. J. G. M. Florax, Peter Nijkamp, Kenneth George Willis (eds.), Comparative Environmental Economic Assessment, Edward Elgar, Cheltenham, UK.

Eshet, T., M.G. Baron, M. Shechter. 2007. Exploring Benefit Transfer: Disamenities of Waste Transfer Stations. Environmental and Resources Economics, 37(3): 521-547.

Groothuis, P.A. 2005. Benefit Transfer: A Comparison of Approaches, Growth and Change, 36 (4): 551-564.

Hanley, N., R.E. Wright, B. Alvarez-Farizo. 2006. Estimating the Economic Value of Improvements in River Ecology Using Choice Experiments: An Application to the Water Framework Directive. Journal of Environmental Management, 78 (2) (2006) 183-193.

H. Jeong, T. Haab. 2004. The Economic Value of Marine Recreational Fishing: Applying Benefit Transfer to Marine Recreational Fisheries Statistics Survey. Working Paper Series (28322), Department of Agricultural, Environmental and Development Economics, Ohio State University, Columbus.

Jiang, Y., S.K. Swallow, M.P. McGonagall. 2005. Context-Sensitive Benefit Transfer using Stated Choice Models: Specification and Convergent Validity for Policy Analysis. Environmental and Resource Economics, 31(4): 477–499.

Johnston, R.J. 2007. Choice Experiments, Site Similarity and Benefits Transfer. Environmental and Resources Economics, 38(3) 331-351.

Johnston, R.J., J. M. Duke. 2009. Willingness to Pay for Land Preservation across States and Jurisdictional Scale: Implications for Benefit Transfer. Land Economics, 85 (2): 217-237.

Kerr, G.N., Basil M. H. Sharp. 2006. Transferring Mitigation Values for Small Stream, in: J. Rolfe, and J. Bennett (eds.), Choice Modeling and the Transfer of Environmental Values, Edward Elgar, Cheltenham, UK.

Kirchhoff, S., B.G. Colby, J.T. LaFrance. 1997. Evaluating the Performance of Benefit Transfer: An Empirical Inquiry. Journal of Environmental Economics and Management, 33 (1): 75-93.

Kristofersson, D., S. Navrud. 2007. Can Use and Non-Use Value be Transferred Across Countries?, in: S. Navrud and R. Ready (eds.), Environmental Value Transfer: Issues and Methods, Springer, Dordrecht, The Netherlands.

Leon, C. J., F. J. Vazquez-Polo, N. Guerre, P. Riera. 2002. A Bayesian Model for Benefit Transfer: Application to National Parks in Spain. Applied Economics, 34 (6): 749-757.

Lindhjem, H., S. Navrud. 2008. How Reliable are Meta-Analyses for International Benefit Transfers? Ecological Economics, 66 (2-3): 425-435.

Loomis, J.B., B. Roach, F. Ward, R. Ready. 1995. Testing Transferability of Recreation Demand Models Across Regions: A Study of Corps of Engineer Reservoirs. Water Resources Research 31 (3): 721–730.

Loomis, J.B. 1992. The Evolution of a More Rigorous Approach to Benefit Transfer: Benefit Function Transfer. Water Resources Research, 28 (3): 701–705.

Matthews, D.I., W.G. Hutchinson, R. Scarpa. 2009. Testing the Stability of the Benefit Transfer Function for Discrete Choice Contingent Valuation Data. Journal of Forest Economics 15 (1-2): 131-146.

Morrison, M., J. Bennett. 2006. Valuing New South Wales Rivers for Use in Benefit Transfer, in: J. Rolfe, and J. Bennett (eds.), Choice Modeling and the Transfer of Environmental Values, Edward Elgar, Cheltenham, UK.

Morrison, M., J. Bennett, R. Blamey, J. Louviere. 2002. Choice Modeling and Tests of Benefit Transfer. American Journal of Agricultural Economics 84 (1): 161-170.

Morrison, M, J. Bennett, R. Blamey, J. Louivere. 2000. Choice Modeling, Non-Use Values and Benefit Transfer. Economic Analysis and Policy, 30 (1): 13-32.

Parsons, G.R., M.J. Kealy. 1994. Benefits Transfer in a Random Utility Model of Recreation. Water Resource Research, 30 (8): 2477-2484.

Piper, S., W.E. Martin. 2001. Evaluating the Accuracy of the Benefit Transfer Method: A Rural Water Supply Application in The USA. Journal of Environmental Management, 63 (3): 223-235.

Ready, R., S. Navrud. 2007. Morbidity Value Transfer, in: S. Navrud and R. Ready (eds.), Environmental Value Transfer: Issues and Methods, Springer, Dordrecht, The Netherlands.

Ready, R., S. Navrud, B. Day, R. Dubourg, F. Machado, S. Mourato, F. Spanninks, M.X.V. Rodriquez. 2004. Benefit Transfer in Europe: How Reliable are Transfers between Countries? Environmental and Resource Economics, 29 (1): 67–82.

Rosenberger, R.S., J.B. Loomis. 2000. Using Meta-Analysis for Benefit Transfer: In-Sample Convergent Validity Tests of an Outdoor Recreation Database. Water Resources Research, 36 (4): 1097-1107.

Rozan, A. 2004. Benefit Transfer: A Comparison of WTP for Air Quality Between France and Germany. Environmental and Resource Economics, 29 (3): 295–306.

Shrestha, R. K., John B. Loomis. 2003. Meta-Analytic Benefit Transfer of Outdoor Recreation Economic Values: Testing Out-Of Sample Convergent Validity. Environmental and Resources Economics, 25 (1): 79-100.

Shrestha, R.K., J.B. Loomis. 2001. Testing a Meta-Analysis Model for Benefit Transfer in International Outdoor Recreation. Ecological Economics 39 (1): 67-83.

Stapler, R. W., Robert J. Johnston. 2009. Meta-Analysis, Benefit Transfer, and Methodological Covariates: Implications for Transfer Error. Environmental and Resources Economics, 42 (2): 227-246.

VandenBerg, T.P., G.L. Poe, J.R. Powell. 2001. Assessing The Accuracy Of Benefits Transfers: Evidence From A Multi-Site Contingent Valuation Study Of Groundwater Quality, in: J.C. Bergstrom, K.J. Boyle, G.L. Poe (eds.), The Economic Value of Water Quality, Edward Elgar, Northampton, MA.

M. Zandersen, M. Termansen, F.S. Jensen. 2007a. Evaluating Approaches to Predict Recreation Values of New Forest Sites. Journal of Forest Economics, 13 (2-3): 103-128.

M. Zandersen, M. Termansen, F.S. Jensen. 2007b. Testing Benefits Transfer of Forest Recreation Values over a Twenty-Year Time Horizon. Land Economics 83 (3): 412-440.

Chapter III. Elicitation Effects in Surveys: A Structural and Simulations Based Analysis

Abstract

Surveys are commonly used to support policy-making and conduct academic research. Inferences based on survey data assume that respondents reveal their true preferences and response data are procedurally invariant to the design and framing of questions. Empirical research indicates that incentive compatibility and procedural invariance does not always hold. Mechanisms have been proposed that explain response anomalies in survey data, however, they are largely identified through ex-post story telling or ex-post econometric adjustments. In addition, we lack a clear understanding on whether these mechanisms are different concepts or similar concepts with different names, or empirically distinguishable. This chapter proposes an alternative structural and simulations based framework to understand how different mechanisms create anomalies in response data. We focus on the contingent-valuation survey format that is used to elicit preferences and willingness to pay for non-market goods. This analysis suggests that some mechanisms are observationally distinguishable and some are empirically equivalent. Additionally, simulations show that specific elicitation effects can explain the observed empirical cumulative density functions of response variables, and that some ex-post story telling is not relevant. These results are useful for considering alternative survey implementation procedures and framing of questions, as well as the development of ex-post adjustments to control for undesirable response effects.

Key Words: elicitation effects, surveys, contingent valuation, dichotomous choice

1. Introduction

Surveys are commonly used to collect economic information by governments, businesses and NGOs to support decision-making and these data are often used to support academic research. Examples include the National Health Interview Survey that monitors health of the U.S. population[19], the National Survey of Fishing, Hunting and Wildlife-Associate Recreation that monitors public use of fish and wildlife resources in the U.S.[20] and Google's Consumer Surveys that collect data from internet users to support market research[21]**.** Designing these and other surveys is a challenging process (Dillman, 2007; Fowler, 2009; Groves et al., 2009). An ideal survey instrument should be procedurally invariant; that is, changes in survey design and the framing of question(s) should be neutral in the elicitation of responses (McCollum and Boyle, 2005; Bateman et al. 2009).

Empirical evidence shows that incentive compatibility and procedural invariance does not always hold. For example, Conti and Pudney (2011) analyze the effect of survey design on reported job satisfaction using the British Household Panel Survey and find that "apparently minor differences in survey design lead to substantial biases in econometric results" (page 1087). Hurd et al. (1998) consider different unfolding brackets to elicit consumption and savings amounts in the 1989 Survey of Consumer Finances and the 1994 Consumer Expenditure Survey. They observe that "in the case of savings, variation in starting values … induces a 100 % difference in estimated median savings" (page 379). They claim that the survey influence arises from anchoring because of which "respondents are influenced by cues contained in the question". These and other violations of procedural invariance are a fundamental concern for

[19] http://www.cdc.gov/nchs/nhis.htm
[20] http://www.census.gov/prod/www/abs/fishing.html
[21] http://www.google.com/insights/consumersurveys/home

any type of survey. Changes in survey designs and ex-post statistical corrections can be used to avoid violations of procedural invariance or to control for the effects in survey responses. However, to successfully implement variations of either approach, it is critical to develop a fundamental understanding of how people respond to survey questions. This knowledge can then be used to adjust survey designs or develop calibration procedures for survey responses.

It is common in surveys, including economic surveys, to manipulate question framing to help respondents reveal economic information (see Hurd et al., 1998, example cited above). Alternatively, it is common in economic surveys to posit some type of stimulus that respondents are asked to consider or evaluate in answering the question(s). This is a logical type of design because many economic questions deal with tradeoffs and information about the alternatives must be posited for survey respondents to answer such questions. For example, List and Gallet (2001) document that dichotomous-choice question, which posit a dollar stimulus for subjects to evaluate, are the most common question format to elicit values in field and laboratory experiments.

In this paper we focus our attention on the dichotomous-choice format of contingent-valuation questions that have been used in nonmarket-valuation surveys since the seminal validity study on this topic by Bishop and Heberlein (1979). In dichotomous-choice questions respondents answer yes or no to a monetary stimulus (bid) for some change in the provision of a good or service (Boyle, 2003). There are variants of this format including single-bounded, one-and-a-half bounds and double-bounded questions (Cooper, Hanemann and Signorello, 2002; Hanemann, Loomis and Kanninen, 1991). We focus specifically on the single-bounded and double bounded variants for two reasons. First, Carson and Groves (2007) argue that single-bounded questions are incentive compatible for truthful revelation of preferences. Other variants, including double-bounded questions, do not satisfy this condition. Second, in practice a number of authors have argued that the magnitude of incentives in all variants of dichotomous choice questions can influence responses, including single-bounded questions, which causes violations of procedural invariance (Boyle, Johnson and McCollum, 1997; Green et al., 1998; Herriges and Shogren, 1996). Focusing on the single-bounded and double-bounded questions provides results that are generalizable to all variants of dichotomous-choice question because the investigations include initial and follow-up bid effects on responses. The other variants of dichotomous-choice questions simply have different variations of follow-up bids.

Numerous labels and explanations have been proposed for anomalies in responses to dichotomous-choice questions. For example, "anchoring" is suggested when bid amounts convey implicit information to survey respondents, perhaps indicating the quality of the item being valued (Boyle, Johnson and McCollum, 1997; Green et al., 1998). "Yea saying" ("nay saying") might occur when respondents will answer yes (no) independent of the magnitude of the monetary incentive (Blamey, Bennett and Morrison, 1999; Holmes and Kramer, 1995). "Social desirability" occurs when respondents answer to please an interviewer, perhaps answering yes to appear as a good citizen (Ethier et al., 2000; Leggett et al., 2003). Framing occurs when respondents frame a higher bid as a loss prospect that increases the likelihood of no responses (DeShazo, 2002).

While these labels have been suggested, no one, to our knowledge, has critically assessed whether these phenomena can be empirically distinguished. Let us consider a specific empirical example to formalize the basic concern. A number of dichotomous-choice studies have been plagued by what is termed the "fat tail" problem where 20% to 40% of respondents continue to answer yes, they would pay, to very high bids (Haab, 1999; Kerr, 2000). Each of the three response anomalies described above could lead to fat tails. If higher bids indicate higher quality, then this "anchoring" could cause respondents to continue to answer yes to higher bids because they perceive they are buying a different (better) good or service. "Yea saying" is also a potential explanation because subjects answer yes independent of bid amounts. Likewise, "social desirability" could also cause people to answer yes to higher bid amounts. This leaves open questions:

1) Are these really different response anomalies or are they different names for the same phenomena?
2) Are these different response anomalies, but observationally equivalent?
3) Are these different response anomalies that interact to form an overall effect?

Thus, one or multiple phenomena can cause fat tails and other anomalies in response data from stated-preference questions. In this paper, we use simulations to develop insights on these responses anomalies to stimulate and guide future theoretical and empirical research.

No one, to our knowledge, has attempted to simultaneously investigate multiple sources of anomalies in responses to dichotomous-choice questions using a structured, utility-theoretic framework. In this paper, we propose a unified utility theoretical framework to explain response anomalies in dichotomous-choice questions and their consequent effects on responses to survey questions. We ask if elicitation effects like anchoring, yea-saying, social desirability etc. are separate phenomena with similar and overlapping manifestations, or separate phenomena and observationally distinguishable. Identifying these effects in field or experimental data is very challenging so we conduct simulations to examine the effect of each anomaly independently on response data.

We find that some phenomena are potentially observationally distinguishable, such as anchoring and yea saying, and these response anomalies can severely alter response distributions to give rise to the fat-tail observations. Further, yea saying and social desirability may be distinct effects, but observationally it is difficult to identify these two distinct effects. On the other hand, multiple phenomena can explain fat tails in response data and yea-saying is unlikely to be the sole reason causing a fat tail. While the results of any simulation critically depend on the assumptions employed, we make our assumptions explicit below and note that our general results are robust to alternative theoretical explanations. These results are useful for considering alternative survey implementation procedures and alternative framing of questions, as well as the development of ex-post adjustments to response data to control for undesirable elicitation effects.

2. Response Anomalies in Surveys: Structural Simulations Analysis

If we assume that everyone holds a non-zero, positive value for a good being valued, then the response distribution of yes/no responses to different bids should look like an inverse cumulative distribution function that asymptotically approaches one (1) as bid values approach $0 from above and asymptotically approaches zero (0) as bid values increase (Figure 1) (Miller et al., 2011). While economic theory does not inform us, exactly, about the shape of the response distribution, logic and economic intuition suggest these conditions should hold. If people hold non-zero positive values, then logic suggest that most people would be willing to pay some amount for the item being valued. At the other end of the distribution, budget constraints and the availability of substitutes suggest that there is an amount beyond which most people would not pay for the item.

The presence of a fat tail implies the right tail of the distribution does not asymptotically approach zero as bid amounts increase (Figure 1), and this has generated discussion of explanations such as anchoring, yea saying and social desirability to explain this observed anomaly in dichotomous-choice question response data. On the other hand, nay saying could result in the response distribution not approaching one at low bid levels (Figure 1). Most of the focus in the literature has been at the upper end of the response distribution because of the concern that estimated values are overstated (List and Gallet, 2001; Little and Berrens, 2004; Murphy et al., 2005). That is, estimated values are computed as the area under the empirical cumulative density function derived from response data and anomalous effects in the upper tail potentially bias welfare estimates upward. Thus, ex-post adjustments have focused on trimming procedures that lower value estimates (Alberini, Kanninen and Carson, 1997; Champ and Bishop, 2001; Chien, Huang and Shaw, 2005; Herriges and Shogren, 1996).

It is important, however, to consider the entire response distribution. Morrison and Brown (2009), for example, report response distributions that are relatively flat, even for an experiment with actual cash transactions (see Figure 1, p. 315). Asymmetric nay saying and yea saying, or anchoring, could cause such a flattening of the response distribution. Further, Morrison and Brown (2009) and Miller et al. (2011) both indicate that response distributions for cash and hypothetical transactions can be very similar. Johnston (2006) and Vossler and Kerkvliet (2003) also show that hypothetical referendum responses (one format of decision rule for a dichotomous-choice question) map to response data from actual referendums. Thus, anomalies in response distributions are not restricted solely to the upper ends of response distributions and may not solely be a characteristic of responses to hypothetical questions. Thus, investigating these types of responses is important in understanding how people answer economic survey questions in general and it is important to consider potential effects throughout response distributions in this analysis.

In the next section we begin with the theoretical structure we use to design our simulations to evaluate response anomalies in dichotomous-choice survey data and then move to discussion of the theoretical frameworks underlay each of our simulations.

2.1. Utility Theoretic Model

Consider a contingent-valuation survey that is aimed at eliciting willingness to pay (*WTP*) for a proposed change in the quality of a non-market good. *WTP* for a change in quality (from q^0 to q^1) is:

$$V_i(y_i - WTP_i, q_i^1, X_i; \alpha_i) = V_i(y_i, q_i^0, X_i; \alpha_i), \qquad (1)$$

where V_i and y_i are utility and income of i^{th} respondent, respectively. α_i and X_i are respondent specific preferences and demographic variables. The closed form solution for *WTP* is:

$$WTP_i = WTP_i(y_i, q_i^0, q_i^1, X_i, \alpha_i). \qquad (2)$$

In a single-bounded, dichotomous-choice question, individuals are asked whether they are willing to pay a stated bid amount. Response of i^{th} individual to first question is:

$$RES_{i1} = \begin{cases} "yes" \; if \; WTP_i \geq BID_{1k} \\ "no" \; if \; WTP_i < BID_{1k} \end{cases}, \qquad (3)$$

where BID_{1k} denotes the k^{th} first bid. In a double-bounded format, an additional question is asked based on the response to first bid. If a respondent says yes (no) to BID_{1k}, a higher (lower) second bid is presented. Response to the m^{th} second bid (BID_{2m}**)** is:

$$RES_{i2} = \begin{cases} "yes" \; if \; WTP_i \geq BID_{2m} \\ "no" \; if \; WTP_i < BID_{2m} \end{cases} . \qquad (4)$$

Empirical evidence indicates the observed responses can be different from the actual responses shown in equations (3) and (4). In Table 1, we summarize common causes of response anomalies in single and double-bounded formats. Below, we discuss these causes and outline the underlying mechanisms that disincentive individuals from revealing truthful responses.

2.1.1. Anchoring

There is an abundance of empirical evidence, which demonstrates that anchoring significantly affects survey responses by distorting quality perception of the item being valued and creating a bias in WTP (Ariely, Lowenstein and Prelec, 2003; Boyle, Johnson and McCollum, 1997; Chien, Huang and Shaw, 2005; Herriges and Shogren, 1996; Green et al., 1998). Anchoring can be uniquely modeled by establishing its relationship with the bid amounts. To do this, we differentiate between two forms of anchoring. First, a bid can exert a contemporaneous effect by influencing quality in the same question in which the bid is presented (Boyle, Johnson and McCollum, 1997). This implies that BID_{1k} (BID_{2m}) influences quality and response in the first (second) question. Second, a bid may exert a lagged effect by influencing a future response (Herriges and Shogren, 1996). For example, in the double-bounded question BID_{1k} can potentially affect quality perception and response to the follow-up question.

To see the effect of anchoring, substitute quality as a function of bid in equation (2) and take the derivate:

$$\frac{\partial WTP_i(y_i,q_i^0,q_i^1(BID),X_i,\alpha_i)}{\partial BID} = \frac{\partial WTP_i(y_i,q_i^0,q_i^1(BID),X_i,\alpha_i)}{\partial q_i^1} \times \frac{\partial q_i^1(BID)}{\partial BID}, \qquad (5)$$

where *BID* represents the first or the second bid. Assuming that individuals value quality positively, the overall effect of bid on *WTP* depends on the sign of $\frac{\partial q_i^1(BID)}{\partial BID}$. If $\frac{\partial q_i^1(BID)}{\partial BID} \lesseqgtr 0$ then $\frac{\partial WTP_i}{\partial BID} \lesseqgtr 0$. Respondents who are likely to say yes to a bid $(WTP_i \geq BID)$ may think that an inferior quality good is being provided because the bid is lower than the value of the item. This can cause a downward bias in quality. In contrast, respondents who value quality less than the bid $(WTP_i < BID)$ may associate the bid with a superior quality good. Thus, anchoring can potentially have asymmetric effects on responses. For our simulations, we put additional structure on $q_i^1(BID)$ in equation (5) to predict the direction of effects of bids on *WTP*.

2.1.2. Yea (Nay) Saying, Warm Glow and Social Desirability

Yea (nay) saying occurs when respondents agree (disagree) with the interviewer regardless of their true preferences or the information provided in the survey that increases the probability of yes (no) responses to dichotomous-choice questions (Boyle et al., 1998; Blamey, Bennett, and Morrison, 1999; Yeung et al., 2006). Using this intuition with equation (1), we have:

$$V_i(y_i - WTP_i^*, q_i^1, X_i, \alpha_i) = V_i(y_i, q_i^0, X_i, \alpha_i), \qquad (6)$$

where $WTP_i^* > WTP_i$ for yea saying and $WTP_i^* < WTP_i$ for nay saying. In equation (6), respondents who indulge in yea saying will overstated their WTP and vice-versa.

We can use equation (6) to model other elicitation effects like warm glow, social desirability and guilt. Warm glow occurs when respondents derive moral, social, and psychological satisfaction by contributing towards a good (Andreoni, 1990; Nunes and Schokkaert, 2003). Social desirability occurs when respondents answer to please an interviewer, perhaps answering yes to appear as a good citizen (Ethier et al., 2000; Leggett et al., 2003).

Additionally, respondents may feel guilty if they do not contribute towards the provision of a good (Bateman et al., 2001). Warm glow, social desirability and guilt will increase the probability of yes responses, which makes them observationally equivalent to yea saying. Thus, even though these mechanisms are separate phenomena we may not be able to identify them separately like we can for anchoring. Similarly, feelings of indignation (because of asking for money), free riding and wastage of resources negatively affect the probability of yes responses that make them observationally equivalent to nay saying (Bateman et al., 2001; Cameron and Quiggin, 1994).

Additionally, the framing effect can be modeled using the nay saying mechanism. In the double-bounded question format, respondents can frame a higher follow-up bid as a loss prospect, which creates a downward bias in *WTP* (DeShazo, 2002). Consider a respondent who agrees to pay the first bid, i.e. $(WTP_i - BID_{1k}) > 0$. In follow-up question, respondent's surplus equals $(WTP_i - BID_{2m})$, which if compared with the first bid $[(WTP_i - BID_{1k}) - (BID_{2m} - BID_{1k})]$ looks like a loss prospect because $(BID_{2m} - BID_{1k}) > 0$. So even if $(WTP_i - BID_{2m}) > 0$, a loss averse respondent may refuse to pay the second higher bid because of loss aversion. Notice that framing only affects respondents who agree to pay the first bid.

2.1.3. *Structural Shifts in Preferences*

Respondents' preferences or *WTP* may change because of being informed or the information asymmetry created by the information provided in the survey (Alberini, Kanninen and Carson's, 1997). For example, a respondent may update the preference for quality due to bids presented in surveys. Lower bids can be suggestive of inferior quality that can potentially reduce the worth of quality in respondent's overall utility. With structural shifts, preference parameters in equation (2) are written as:

$$\alpha_i = \alpha + u_i, \tag{7}$$

where α is a systematic shift (positive or negative) and u_i is a random error component. With preferences instability the WTP function for second question is:

$$WTP_i = WTP_i(y_i, q_i^0, q_i^1\ , X_i, \alpha, u_i). \tag{8}$$

With this set up, no a priori predictions can be made for preferences instability because of the random component (u_i). However, if we parameterize α_i as function of survey information (e.g. bid) and assume that $E(u_i) = 0$, we have:

$$WTP_i = WTP_i\big(y_i, q^0, q^1,\ \alpha_i^*, X_i\big), \tag{9}$$

where $\alpha_i^* = h(BID)$. Similar to quality specific anchoring, structural shifts due to bid will affect *WTP* depending on the sign of $\frac{\partial h(BID)}{\partial BID}$. If the bid increases the marginal utility of quality, respondents will overstate their *WTP* and vice-versa.

In our next section, we put additional structure on our general model and empirically investigate the effect of each mechanism while keeping the effect of other mechanisms constant.

2.2. Simulations

We start by assuming a linear utility function with heterogeneous preference parameters:

$$V_i = \alpha_{iy} y_i + \alpha_{iq} q + \alpha_{ix} X_i + \alpha_{iqx} X_i q, \tag{10}$$

which is used to solve the WTP function as:

$$WTP_i = \frac{(\alpha_{iq}+\alpha_{iqx}X_i)\Delta q}{\alpha_{iy}}, \tag{11}$$

where $\Delta q = q^1 - q^0$. To examine the impact of elicitations effects, we first create a true simulated WTP function and then incorporate different mechanisms.

The simulated true WTP density function is created by using data on X_i (see Table 2) and bids from an actual contingent valuation survey that was conducted to elicit WTP for reducing eutrophication in rivers and lakes (Bateman et al. 2009). In this survey eight first bids were presented to all respondents. We drop missing data and observation with BID_{1k} =\$48.5. Additionally, we trim the Bateman et al. data for our simulations so that each first bid is presented to equal number of respondents. The minimum number of observations corresponding to a first bid is 81 and the maximum is 419. So, we keep the first 81 observations associated with each bid, which results in 567 observations for a total of 7 bids. The second bids are endogenous to our simulations and generated depending the minimum and maximum of *WTP*. Additionally, the second bid is assigned to respondents based on their first responses. Respondents on the ascending sequence (who say yes to the first bid) are assigned a higher bid and respondents on the descending sequence (who say no to the first bid) are assigned a lower bid.

The preference parameters are assigned such that respondents have heterogeneous WTP. Specifically, for equation (11) we assume that $\alpha_{iq}, \alpha_{iqx} \sim Uniform(min, max)$ for all *i* and $\alpha_{iy} = 0.4$, and normalize q^0 and Δq to 1. For different combinations of (*min, max*) we take 2000 draws of $\alpha_{iq}, \alpha_{iqk}$ and compute WTP for 567 respondents across all draws. Mean of WTP across 2000 draws for 567 respondents is used to estimate the WTP density function. The density functions are estimated using Gaussian Kernel and Silverman’s rule of thumb bandwidth.[22] Figure 2a shows the density plots of mean WTP for 567 respondents estimated via equation (11) using varying combinations of *min* and *max*.

We choose WTP function with $\alpha_{iq}, \alpha_{iqx} \sim U(0.2,0.7)$ as our true function (see Figure 2a). The parameters for the simulated true WTP are chosen such that the mean of WTP (of all respondents) lies in the range of WTP estimates ($76 - $101) presented by Bateman et al. (2009). Additionally, we wanted the resulting response functions (average yes response at each bid) to be well-behaved. The true WTP and response probability functions are shown in Figure 2b. The mean and the standard deviation of WTP are $100 and 91 respectively. Response functions are drawn separately for the first bid, ascending and descending sequences. The second bid responses are dependent on the first bid response. For example, 100% of respondents who are offered the first bid of $10 agree to pay it and they are offered $20 in the second question (ascending sequence). Thus, we have no observations for the $5 bid on the descending sequence. Next, we incorporate elicitation effects to examine the direction of effects on WTP and response functions.

2.2.1. Anchoring

The updated quality in the *j*th question (q_i^j) can potentially decrease or increase when compared to the actual quality (q^1) because of anchoring. We model q_i^j such that respondents across ascending sequence ($WTP_i \geq BID_j$) place lower values on quality and vice-versa:

$$q_i^j(BID_j) = (1-\gamma_i)q^1 + \gamma_i q_{i,BID_j}, \text{ where} \tag{12}$$

[22] The rule of thumb bandwidth is estimated as $\hat{h} = \hat{\sigma} n^{-1/5}$, where $\hat{\sigma}$ is the sample standard deviation and *n* is the sample size.

$$q_{i,BID_j} = \begin{cases} (1-\theta_i) * q^1 * I(WTP_i \geq BID_j) \\ (1+\theta_i) * q^1 * I(WTP_i < BID_j) \end{cases}, \qquad (13)$$

where BID_j is the first or the second bid. In equation (12), γ_i is the respondent specific anchoring parameter with $0 \leq \gamma_i \leq 1$. If $\gamma_i = 1$, we have complete anchoring and $\gamma_i = 0$ shows no anchoring effect on quality. Quality indicated by the bid in the j^{th} question is given by q_{i,BID_j} in equation (13) where $I(.)$ equals "1" iff the condition in the parenthesis holds. Additionally, $0 \leq \theta_i \leq 1$ and $(1-\theta_i) * 100\%$ measures the percent change in quality. Impact of anchoring is severe when $|\theta_i| \rightarrow 1$ and $\gamma_i \rightarrow 1$.

We substitute equations (12) and (13) in equation (11) and investigate multiple scenarios. We examine changes in WTP and response probability functions with full sample (100% of responses are affected by the bids) and partial sample (50% of the responses are affected by the bids) anchoring. For each of these scenarios, we randomly assign high values of γ_i, θ_i (0.5 to 1) to *p*% of respondents and low values of γ_i, θ_i (0 to 0.5) to the remaining *100-p*% of respondents in 2000 draws. WTP of true model and models with the bid effect are tested using t-test for the equality of means, F-test for the equation of variances and Mann-Whitney (MW) test for equality of medians. The percent of significant *p*- values in 2000 draws are reported for each test and the significance level is set at 95%.

In Table 3, consider the second row that examines the effect of first bid on the mean, the standard deviation and the median of WTP (of all the respondents across draws) when 90% of all respondents are assigned low values of γ_i, θ_i and the remaining 10% are assigned high values of γ_i, θ_i. The mean and the median of observed WTP as compared to true WTP decrease by about 7% and 2%, respectively. These differences as shown by the percent of significant t-test (column 3) and MW-test (column 7) are not significant. However, the first bid has a much severe effect on the spread of WTP as the standard deviation drops by about 18% and the difference is significant, we reject the null hypothesis of equal variances in 99% of 2000 draws. We also consider the second bid effect assuming that the first bid effect does not exist. Qualitatively, the second bid has the same effect on WTP function (mean, median and standard deviation decrease) as the first bid.

A common finding for first and second bid anchoring is that on increasing the percent of respondents with high values of γ_i, θ_i, and the mean, the median and the standard deviation decrease significantly. This finding can be explained by substituting equations (12) and (13) in equation (11) and solving for the difference between the observed WTP and actual WTP:

$$WTP_i^o - WTP_i = \begin{cases} -k * I(WTP_i \geq BID_j) \\ k * I(WTP_i < BID_j) \end{cases} \tag{14}$$

where $k = \frac{(\alpha_{iq}+\alpha_{iqx}X_i)\gamma_i\theta_i}{\alpha_{iy}} > 0$, WTP_i^o is the observed WTP and WTP_i is the actual WTP specified in equation (11). The observed WTP is lower than actual WTP across the ascending sequence and reverse holds for descending sequence. Thus, the net effect on WTP depends on difference between the proportions of respondents who indulge in anchoring across the ascending and the descending sequences. In our true model, about 53% of the respondents have $WTP_i \geq BID_{1k}$. The bids cause reductions in $q^1$for these respondents so the net bid effect is a leftward shift in WTP functions (Figure 3a.).

Anchoring due to the first bid can exert influence on the first question response (contemporaneous anchoring) and the second question response (lagged anchoring). Contemporaneous anchoring causes flatness in the response probability functions as less (more) respondents agree to pay lower (higher) bids (Figure 3a. and Figure 3c.). In contrast, lagged anchoring works like nay (yea) saying across the ascending (descending) sequence (Figure 3b). The scenarios with partial sample when 50% of respondents are affected by anchoring are also examined in Table 3 and Figure 4. As expected, the impacts on WTP and response probability functions are less severe. For example, with 90% low anchoring, equality of variances is rejected in only 31% of draws (as compared to 100% in case of full sample anchoring) and response probability curves shift marginally.

2.2.2. Yea (Nay) Saying

For tractability, we assume that WTP is additively separable in yea (nay) saying. Using equations (6) and (11), with a separable yea (nay) saying effect the observed WTP is:

$$WTP_i^o = WTP_i + \delta_i. \qquad (15)$$

where δ_i signifies the vector of individual specific yea (nay) saying parameters. Yea saying causes overstatement of *WTP* irrespective of the monetary stimuli ($\delta_i > 0$) and nay saying decreases the background probability of yes responses ($\delta_i < 0$). With $\delta_i = 0$ respondents will reveal their true preferences.

We assume that $\delta_i \sim \ln \mathrm{N}\,(u, \sigma)$.[23] With varying combinations of u and σ, we ask what happens when 100% (full sample) or 50% (partial sample) of the respondents indulge in yea saying? Table 4 and Figure 5 present these results. Consider the second row. With small values for mean and s.d. for δ_i yea saying does not have a significant effect on WTP. However, as we increase the mean and s.d. of δ_i, WTP density functions shift rightward significantly. The

[23] Changing the assumption about the distribution of δ_i does not change the qualitative inferences of our model. However, it can have implications for the shape of the response probability curves.

response probability curves become flat and when $\delta_i \sim \ln \text{N}\,(3{,}2)$ the tails of response probability curves increase by almost 20% (more than 20% for descending sequence). Partial yea – saying has a similar but less severe effect (Figure 6).

Opposite inferences are drawn in case of nay saying, which is modeled with a negative shift effect (Table 5 and Figure 7a.). The WTP functions shift leftward that cause downward shifts in the response probability functions. Effects like framing and indignation can be modeled in a similar fashion. We present the outcomes with framing effect in Table 5 and Figure 7b.

2.2.3. Structural Shifts in Preference

Assuming that bids may change the marginal utility of quality, we can model the updated preference parameter as:

$$\alpha_{iq}^{*} = \begin{cases} (1-\vartheta) * \alpha_{iq} * I(WTP_i \geq BID_j) \\ (1+\vartheta) * \alpha_{iq} * I(WTP_i < BID_j)' \end{cases} \qquad (16)$$

where α_{iq} are the true preferences. For our simulations, we fix $\vartheta = (0.1{,}0.2{,}0.3{,}0.4)$. By formulation, the effect of structural shift is same as that of anchoring. We assume that a higher bid in comparison to respondent's WTP may perhaps indicate the importance of quality change. Table 5 and Figure 7c show that preference shifts can also cause leftward shifts in WTP functions and flatness in the response probability functions.

2.3. Robustness Check

We also examine the inferences of our theoretic framework using a more flexible Constant Elasticity of Substitution utility function. The utility function is specified as:

$$V_i = \left[q_i^{\delta_i} + X_i^{\delta_i} + y_i^{\delta_i}\right]^{\frac{1}{\delta_i}}, \tag{17}$$

where y_i represents respondent specific income and $0 \neq \delta_i \leq 1$. Using equations (1) and (17), the solution for WTP is:

$$WTP_i = y_i - \left[q_i^{0\delta_i} - q_i^{1\delta_i} + y^{\delta_i}\right]^{\frac{1}{\delta_i}}. \tag{18}$$

Assuming that $\delta_i = U(min, max)$, we vary min and max such and plot WTP density functions (Figure 8a). We choose the simulated truth such that mean of WTP is similar to that of the linear utility case and the response probability functions are well behaved. Using this, we pick $\delta_i = U(0.2{,}0.36)$ to represent the true model. Table 6 and Figures 8b and 8c examine the effects of first bid contemporaneous anchoring and yea saying on WTP density and response probability functions. Anchoring causes reduction in the mean and standard deviation of WTP whereas yea saying shifts the WTP function rightwards. However, the magnitude of effects may be different. For example, with 90% low anchoring equality of means of true and anchored WTP is rejected in 32% of draws (in case of linear utility it was 0.2%). Thus, qualitatively our results remain the same irrespective of the functional form for utility.

3. Discussion

With our structural simulations analysis, we demonstrate that elicitation effects can play a major role in distorting inferences from dichotomous-choice survey questions. Several effects can have confounded effects and some elicitation effects can be theoretically and empirically distinguished. Below, we discuss our main findings.

3.1 Contemporaneous anchoring causes flatness in the response probability curves. Even when a small percent of respondents indulge in contemporaneous anchoring, WTP function is significantly affected.

3.2 Lagged anchoring in the second question operates like yea / nay saying. On the ascending sequence, respondents will indulge in nay saying whereas respondents on descending sequence are more likely to say yes. Lagged anchoring also causes flatness in the response probability functions. However, its effect is less severe when compared to contemporaneous anchoring.

3.3 Yea (nay) saying shifts the response probability curves upward (downward). However, the shift depends on how respondents resort to yea (nay) saying. For example, with full sample yea saying the tails of the response probability function are likely to shift. However, if the yea saying parameter is distributed non – uniformly or there is incomplete yea saying, the results can vary.

3.4 By definition, elicitation effects like warm glow, social desirability and guilt are observationally equivalent to yea saying. In contrast, indignation, framing and free riding will be observationally indistinguishable from nay saying.

3.5 If respondents' preferences shift due to information provided in the surveys, it can potentially create bias in response data. For example, the assigned bid may affect the marginal utility of the quality change or the good. This can also cause flatness in the response probability curve.

3.6 A fat tail in response data can be explained by multiple mechanisms. If the empirical cumulative density function resembles Figure 1, it is potentially due to anchoring that increases the likelihood of no responses at lower bids and vice-versa. However, if the empirical cumulative density function exhibits a fat tail at higher bids only, it can potentially be caused by yea saying or other equivalent mechanisms like warm glow or social desirability.

4. Conclusions

Survey data should be procedurally invariant to the framing of the questions for building credible inferences. However, empirical research shows that response anomalies occur in survey data. Several mechanisms have been proposed that explain these response anomalies and survey designs and ex-post econometric corrections have been suggested to mitigate unintended consequences of such mechanisms. However, prior to these corrections, it is important to understand whether we can empirically distinguish between these survey effects.

To do this, we explore the contingent-valuation survey format that is commonly used in the economics literature to elicit stated preferences for different transportation outcomes (Hultkrantz, Lindberg and Andersson, 2006), different health outcomes (Ryan and Watson, 2009), and changes in environmental quality (Hanley et al., 1998). We propose a structural utility model to understand how individuals respond to surveys. We then incorporate the potential mechanisms that explain the divergence in actual and observed responses. We conduct statistical simulations to empirically investigate the empirical outcomes of our theoretical model. While we consider response anomalies in stated-preference data, these anomalies do occur in other types of surveys (Bachman and O'Malley, 1984; Conti and Pudney, 2011; Ethier et al., 2000; Fisher, 1993; Hurd, 1999; Hurd et al., 1998).

A number of mechanisms that cause response anomalies are observationally equivalent. For example, elicitation effects like yea saying, warm glow, guilt, social desirability have equivalent effects on willingness to pay function. Similarly, empirically we cannot distinguish between nay saying, free riding and indignation. Even though econometric corrections for these confounded effects should reduce the bias, however, it is hard to separate out the effects and explain the cause(s). We find that anchoring and yea saying have severe effects on individual responses. These can potentially explain observed anomalies in response probability functions.

Deeper understanding of response anomalies in survey data can help in devising procedures that mitigate elicitation effects. For example, public polls and environmental based survey research on public goods can potentially be affected by yea saying that causes significant overstatement of values. Further, they are equally likely to be affected by other effects like anchoring that can potentially accentuate the effect of yea saying. Thus, controlling for yea saying alone is unlikely to correct for other mechanisms like anchoring that also causes fat tails in response data. Thus, more research is required on the optimal survey instruments, bid designs and unified ex-post econometric corrections to mitigate the influence of elicitation effects.

References

Alberini, Anna, Barbara Kanninen, and Richard T. Carson. 1997. Modeling Response Incentive Effects in Dichotomous Choice Contingent Valuation Data. Land Economics, 73(3): 309-24.

Ariely, Dan, George Loewenstein, and Drazen Prelec. 2003. Coherent Arbitrariness: Stable Demand Curves Without Stable Preferences. The Quarterly Journal of Economics, 118 (1): 73–105.

Andreoni, James. 1990. Impure Altruism and Donations to Public Goods: A Theory Of Warm-Glow Giving. The Economic Journal, 100: 464–477.

Bachman, J.G., and P.M. O'Malley. 1984. Yea-saying, Nay-saying, and Going to Extremes: Black-White Differences in Response Styles. Public Opinion Quarterly, 48 (2): 491-509.

Bateman, Ian J., Brett H. Day, Daine P. Dupont, and Stavros Georgiou. 2009. Procedural Invariance Testing of the One - and - One Half Bound Dichotomous Choice Elicitation Method. The Review of Economics and Statistics, 91(4): 806-820.

Bateman, Ian J., Ian H. Langford, Andrew P. Jones, and Geoffrey N. Kerr. 2001. Bound and Path Effects in Double and Triple Bounded Dichotomous Choice Contingent Valuation. Resource and Energy Economics, 23(3): 191-213.

Bishop, R.C., and T.A. Heberlein. 1979. Measured Values of Extramarket Goods: Are Indirect Measures Biased? American Journal of Agricultural Economics, 61 (5): 926-930.

Blamey, R.K., J.W. Bennett and M.D. Morrison. 1999. Yea-saying in Contingent Valuation Surveys. Land Economics, 75 (1): 126-141.

Boyle, K.J. 2003. Contingent Valuation in Practice. Chapter 4 in A Primer on Nonmarket Valuation, P.A. Champ, K.J. Boyle and T.C. Brown (eds). Springer.

Boyle, Kevin J., Hugh F. MacDonald, Hsiang-tai Cheng, and Daniel W. McCollum. 1998. Bid Design and Yea Saying in Single Bounded, Dichotomous-Choice Questions. Land Economics, 74(1): 49-64.

Boyle, Kevin J., F. Reed Johnson, and Daniel W. McCollum. 1997. Anchoring and Adjustment in Single-Bounded, Contingent-Valuation Questions. American Journal of Agricultural Economics, 79(5): 1495-500.

Cameron, T.A. and J. Quiggin. 1994. Estimation Using Contingent Valuation Data from a Dichotomous Choice with Follow-Up Questionnaire. Journal of Environmental Economics and Management, 27: 218-234.

Carson, Richard T., and Theodore Groves. 2007. Incentive and Informational Properties of Preference Questions. Environmental and Resource Economics, 37(1): 181-210.

Champ, P.A., and R.C. Bishop. 2001. Donation Payment Mechanisms and Contingent Valuation: An Empirical Study of Hypothetical Bias. Environmental and Resource Economics, 19 (4): 383-402.

Chien, Y., C. J. Huang, and D. Shaw. 2005. A General Model of Starting Point Bias in Double-Bounded Dichotomous Contingent Valuation Surveys. Journal of Environmental Economics and Management, 50(2): 362-377.

Conti, G., and S. Pudney. 2011. Survey design and the Analysis of Satisfaction. The Review of Economics and Statistics, 93 (3): 1087-1093.

Cooper, J.A., M. Hanemannn, and G. Signorello. 2002. One-and-One-Half-Bound Dichotomous Choice Contingent Valuation. The Review of Economics and Statistics, 84 (4): 742-750.

DeShazo, J.R. 2002. Designing Transactions without Framing Effects in Iterative Question Formats. Journal of Environmental Economics and Management, 43(3): 360-85.

Dillman, D.A. 2007. Mail and Internet Surveys: The tailored design method. John Wiley & Sons, Inc.: Hoboken, NJ.

Ethier, R.G., G.L. Poe, W.D. Schulze and J. Clark. 2000. A comparison of Hypothetical Phone and Mail Contingent Valuation Responses for Green Pricing Electricity Programs. Land Economics, 76 (1): 54-67.

Fisher, R.J. 1993. Social Desirability Bias and the Validity of Indirect Questioning. Journal of Consumer Research, 20 (2): 303-315.

Fisher, W.L., A.E. Grambsch, D.L. Eisenhower and D.R. Morganstein. 1991. Length of Recall Period and Accuracy of Estimates from the National Survey of Fishing, Hunting, and Wildlife-Associated Recreation. American Fisheries Society Symposium, 12: 367-374.

Fowler, F.J. 2009. Survey Research Methods. SAGE Publications, Inc.: Thousand Oaks, CA.

Green, D., K.E. Jacowitz, D. Kahneman and D. McFadden. 1998. Referendum Contingent Valuation, Anchoring, and Willingness to pay for Public Goods. Resource and Energy Economics, 20 (2): 85-116.

Groves, R.M., F.J. Fowler, M.P. Couper, J.M. Lepkowski, E. Singer and R. Tourangeau. 2009. Survey Methodology. John Wiley & Sons, Inc.: Hoboken, NJ.

Haab, T.C. 1999. Nonparticipation or Misspecification? The Impacts of Nonparticipation on Dichotomous Choice Contingent Valuation. Environmental and Resource Economics, 14 (4): 443-461.

Habb, T.C., and K.E. McConnell. 1997. Referendum Models and Negative Willingness to Pay: Alternative Solutions. Journal of Environmental Economics and Management, 32 (2): 251-270.

Hanemann, M., J. Loomis and B. Kanninen. 1991. Statistical Efficiency of Double-Bounded Dichotomous Choice Contingent Valuation. American Journal of Agricultural Economics, 73 (4): 1255-1263.

Hanley, N., D. MacMillan, R.E. Wright, C. Bullock, I. Simpson, D. Parsisson and B. Crabtree. 1998. Contingent Valuation Versus Choice Experiments: Estimating the benefits of Environmental Sensitive Areas in Scotland. Journal of Agricultural Economics, 49 (1): 1-15.

Herriges, Joseph A. and Jason F. Shogren. 1996. Starting Point Bias in Dichotomous Choice Valuation with Follow-Up Questioning. Journal of Environmental Economics and Management, 30(1): 112-31.

Holmes, T.P., and R.A. Kramer. 1995. An Independent Sample test of Yea-saying and Starting Point Bias in Dichotomous-choice Contingent Valuation. Journal of Environmental Economics and Management, 29 (1): 121-132.

Hultkrantz, L., G. Lindberg and C. Andersson. 2006. The Value Of Improved Road Safety. Journal of Risk and Uncertainty, 32 (2): 151-170.

Hurd, M.D. 1999. Anchoring and Acquiescence Bias in Measuring Assets in Household Surveys. Journal of Risk and Uncertainty, 19 (1): 111-136.

Hurd, M.D., D. McFadden, H. Chand, L. Gan, A. Menill and M. Roberts. 1998. Consumption and Savings Balances of the Elderly: Experimental Evidence on Survey Response Bias. Chapter 8 in Frontiers in the Economics of Aging, D.A. Wise (ed.). University of Chicago Press.

Johnston, R.J. 2006. Is Hypothetical Bias Universal? Validating Contingent Valuation Responses Using a Binding Public Referendum. Journal of Environmental Economics and Management, 52 91): 469-481.

Kerr, G.N. 2000. Dichotomous Choice Contingent Valuation Probability Distributions. Australian Journal of Agricultural and Resource Economics, 44 (2): 233-252.

Leggett, C.G., N.S. Kleckner, K.J. Boyle, J.W. Duffield and R.C. Cameron. 2003. Social Desirability Bias in Contingent Valuation Surveys Administered Through In-person Interviews. Land Economics, 79 (4): 561-575.

List, J.A., and C.A. Gallet. 2001. What Experimental Protocol Influence Disparities Between Actual and Hypothetical Stated Values? Environmental and Resource Economics, 20 (3): 241-254.

Little, J., and R. Berrens. 2004. Explaining Disparities between Actual and Hypotetical Stated Values: Further Investigation Using Meta-Analysis. Economics Bulletin, 3 (6): 1-13.

Mccollum, D.W. and K.J. Boyle. 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): 23-33.

Miller, K.M., R. Hofstetter, H. Krohmer and Z.J. Zhang. 2011. How Should Consumers' Willingness to Pay Be Measured? An Empirical Comparison of State-of-the-Art Approaches. Journal of Marketing Research, XLVIII (Feb): 172-184.

Morrison, M., and T.C. Brown. 2009. "Testing the Effectiveness of certainty Scales, Cheap Talk, and Dissonance-Minimization in reducing Hypothetical Bias in Contingent Valuation Studies." Environmental and Resource Economics, 44 (3): 307-326.

Murphy, J.J., P.G. Allen, T.H. Stevens and D. Weatherhead. 2005. "A Meta-Analysis of Hypothetical Bias in Stated Preference Valuation." Environmental and Resource Economics, 30 (3): 313-325.

Nunes, Paulo A.L.D., Erik Schokkaert. 2003. Identifying the Warm Glow Effect in Contingent Valuation. Journal of Environmental Economics and Management, 45 (2): 231–245.

Ryan, M., and V. Watson. 2009. Comparing Welfare Estimates from Payment Card Contingent Valuation and Discrete Choice Experiments. Health Economics, 18 (4): 389-401.

Vossler, C.A., and J. Kerkvliet. 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, 46 (3): 631-649.

Table 1: Elicitation Effects in Contingent-Valuation, Dichotomous-Choice Questions*

Elicitation Effect	Mechanism	Format	WTP
Contemporaneous anchoring	$Quality = f(BID_T)$	SB, DB	Bias toward BID_T
Lagged anchoring	$Quality = f(BID_{T-1})$	DB	Bias toward BID_{T-1}
Yea saying, guilt, warm glow, and response acquiescence	$Utility = Utility^0 + \delta_y$	SB, DB	Bias upward
Nay saying, free riding, and indignation	$Utility = Utility^0 - \delta_n$	SB, DB	Bias downward
Framing	$Utility = Utility^0 - \delta_f$	DB Ascending	Bias downward
Preference instability	$Utility = U(\alpha \pm \epsilon)$	SB, DB	Indeterminate

* This table presents various mechanisms that effect willingness to pay and responses in single bounded (SB) and double bounded (DB) formats. Please refer to the text for a detailed description of these mechanisms.

Table 2: Descriptive Statistics of Variables Used in Simulations*

Variable	Definition	Mean (S.D.)
INCOME	Household income per capita per month	\$883.36 (632.50)
AGE	Respondent's age	45.39 (17.61)
SEX	'1' if respondent is a female, '0' if male	0.42 (0.50)
VISITS	Number of visits to local lakes and rivers per year	41.25 (78.06)
USE	'1' if visits to lakes and rivers involving going in or on water, '0' otherwise	0.29 (0.45)
FISH	'1' if visits to lakes and rivers for recreational fishing, '0' otherwise.	0.19 (0.39)
KIDS	Number of children under 16 years of age	0.53 (0.93)
N	Total Observations	567

* This table presents descriptive statistics of the variables used for simulating willingness to pay and responses to SB and DB formats. S.D. refers to the standard deviation.

Table 3: Tests of Equality of WTP Density Functions with Anchoring

Scenario		Mean	t - test	S.D.	F - test	Median	MW- test
			First Bid				
Full Sample	90% low	93.25	0.20%	73.36	99.25%	72.93	0%
	70% low	88.03	80.60%	62.02	99.95%	70.15	52.40%
	50% low	81.55	99.90%	51.36	98.70%	67.92	99.45%
	30% low	73.47	100%	47.99	93.15%	63.06	100%
Partial Sample	90% low	96.69	0%	81.85	31.45%	75.29	0%
	70% low	93.90	0.95%	74.86	44.50%	73.35	1.05%
	50% low	90.37	31.70%	66.86	34%	71.44	31.60%
	30% low	86.01	83.95%	58.07	15.10%	69.70	83%
			Second Bid				
Full Sample	90% low	95.73	0%	80.47	44.9%	74.16	0.10%
	70% low	92.29	8.50%	74.88	23.0%	70.83	61.25%
	50% low	87.78	59.10%	71.27	31.4%	66.86	99.15%
	30% low	82.32	91.80%	73.87	83.5%	65.33	100%
Partial Sample	90% low	97.96	0%	85.35	5.10%	75.37	0%
	70% low	96.29	0.10%	81.68	7.30%	75.02	0.65%
	50% low	94.09	2%	77.36	15.95%	71.74	29.85%
	30% low	91.17	18.15%	73.46	51.35%	69.31	80.25%

Table 4: Tests of Equality of WTP Density Functions with Yea Saying

Scenario		Mean	t- test	S.D.	F - test	Median	MW - test
Full Sample	$\ln \mathcal{N}(0,1)$	101.87	0%	90.96	0%	76.92	0%
	$\ln \mathcal{N}(1,2)$	120.55	98.50%	91.01	93.2%	96.36	100%
	$\ln \mathcal{N}(2,2)$	154.31	99.05%	90.81	100%	129.60	100%
	$\ln \mathcal{N}(3,2)$	247.40	99.05%	94.01	100%	223.75	100%
Partial Sample	$\ln \mathcal{N}(0,1)$	101.01	0%	91.01	0%	75.90	0%
	$\ln \mathcal{N}(1,2)$	110.16	6%	91.26	58.95%	85.71	58.25%
	$\ln \mathcal{N}(2,2)$	127.29	96.95%	91.23	99.9%	103.06	100%
	$\ln \mathcal{N}(3,2)$	174.29	100%	92.63	100%	151.48	100%

Table 5: Tests of Equality of WTP Density Functions with Other Elicitation Effects

Scenario		Mean	t - test	S.D.	F - test	Median	MW - test
			Nay Saying				
Full Sample	ln $\mathcal{N}(0,1)$	98.64	0%	91.19	0%	73.47	0%
	ln $\mathcal{N}(1,2)$	80.16	98%	91.23	93%	55.89	100%
	ln $\mathcal{N}(2,2)$	45.55	98.7%	91.30	100%	21.90	100%
			Framing				
Full Sample	ln $\mathcal{N}(0,1)$	99.61	0%	90.83	0%	74.69	0%
	ln $\mathcal{N}(1,2)$	92.63	1.9%	88.92	31.45%	69.53	4.4%
	ln $\mathcal{N}(2,2)$	79.86	82.9%	87.34	85.95%	58.68	98.9%
			Preference Uncertainty and Anchoring				
Full Sample	20%	92.36	8.25%	71.30	95.4%	73.30	17.8%
	30%	88.37	97.35%	62.57	100%	70.23	91.%
	40%	84.45	100%	55.44	100%	69.21	100%
	50%	80.50	100%	50.20	100%	67.59	100%

Table 6: Robustness Check w/ Constant Elasticity of Substitution Utility Function

Scenario		**Mean**	**t - test**	**S.D.**	**F - test**	**Median**	**MW – test**
				First Bid Anchoring			
Full Sample	90% low	91.99	31.75%	46.52	60.30%	88.35	1.50%
	70% low	85.19	99.95%	41.80	5.65%	83.80	99.85%
	50% low	80.80	100%	39.98	12.95%	76.09	100%
	30% low	75.81	100%	39.04	56.40%	68.41	100%
				Yea Saying			
Full Sample	$\ln \mathcal{N}(0,1)$	100.17	0%	53.77	0.0%	93.06	0%
	$\ln \mathcal{N}(1,2)$	118.84	98.95%	54.07	99.9%	112.04	100%
	$\ln \mathcal{N}(2,2)$	152.55	99.05%	54.70	100%	148.93	100%
	$\ln \mathcal{N}(3,2)$	247.89	99.25%	59.70	100%	243.96	100%

Figure 1: Inverse Cumulative Density Function of Response Data

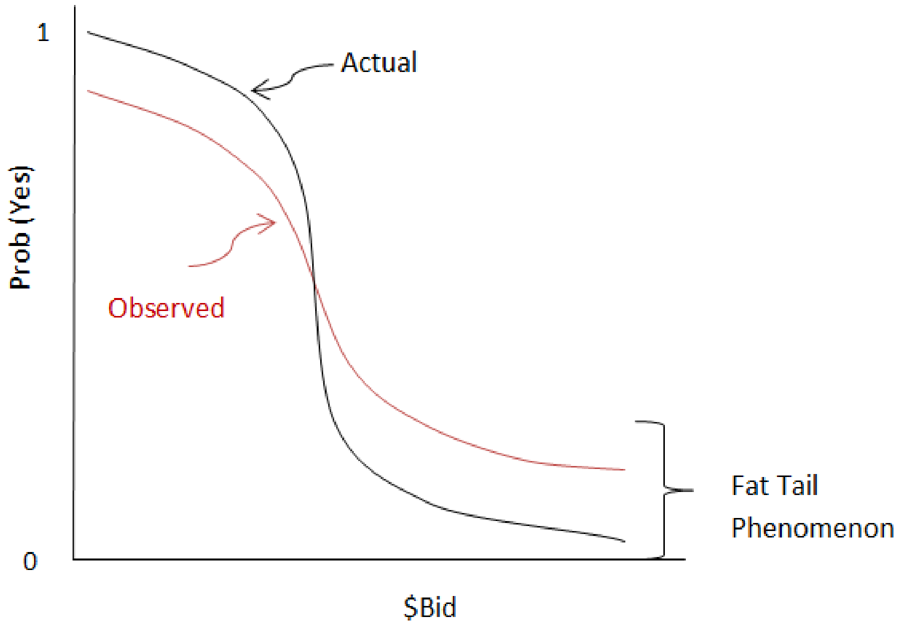
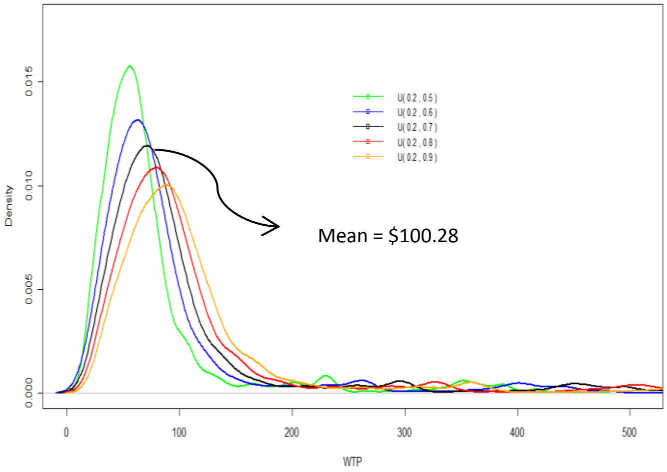


Figure 2: Simulated Willingness to Pay (WTP) Density Functions

Figure 2a: WTP Density Functions w/ Varying *min* and *max* Values*



*Density functions are plotted using Gaussian Kernel and Silverman’s rule of thumb bandwidth.

Figure 2b: WTP Density and Response Probability Functions-Simulated Truth*

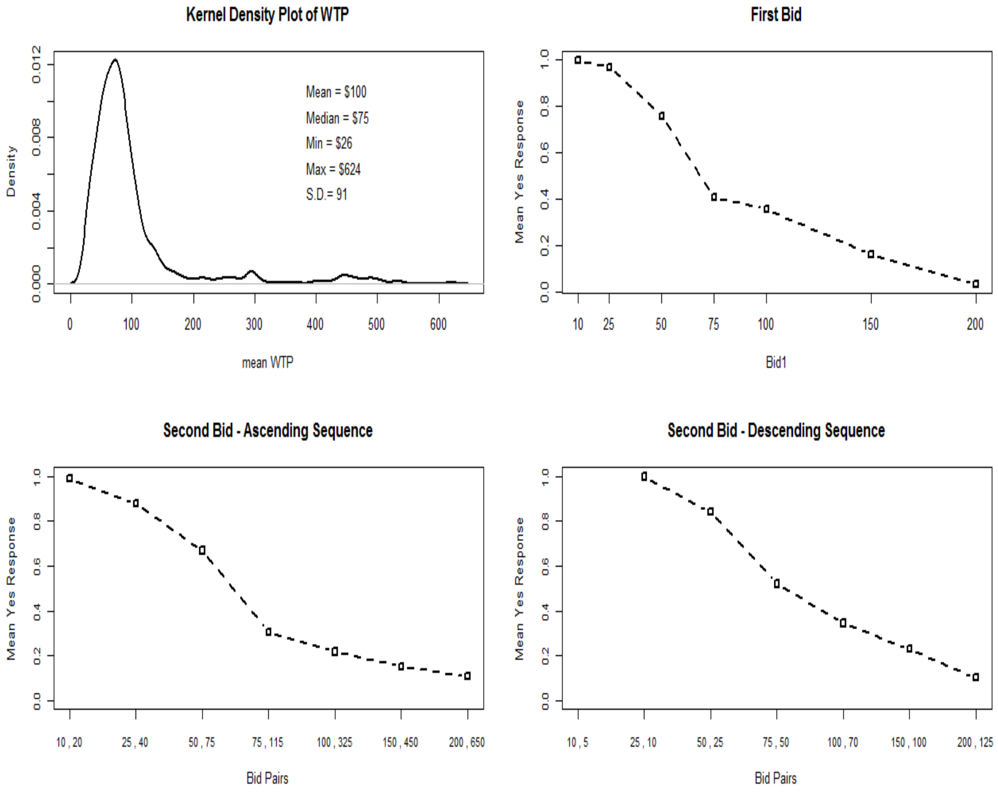


Figure 3: Full Sample Anchoring

Figure 3a: Contemporaneous Anchoring: First Bid

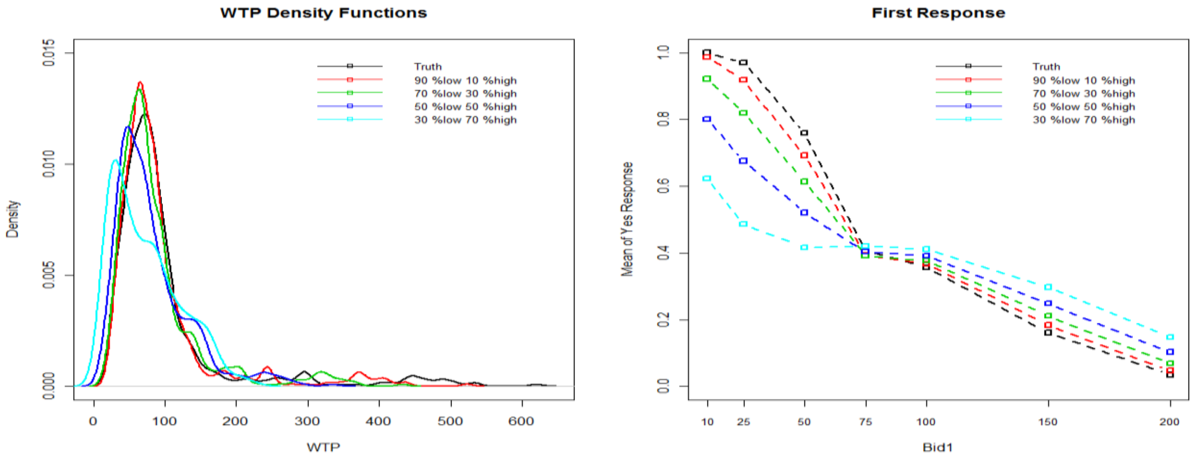


Figure 3b: Lagged Anchoring: First Bid

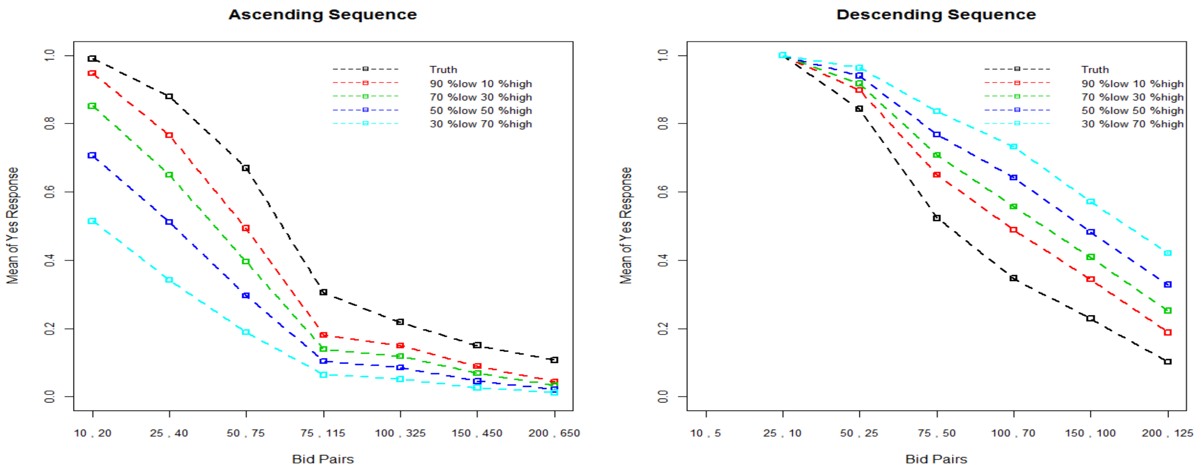


Figure 3c: Contemporaneous Anchoring: Second Bid

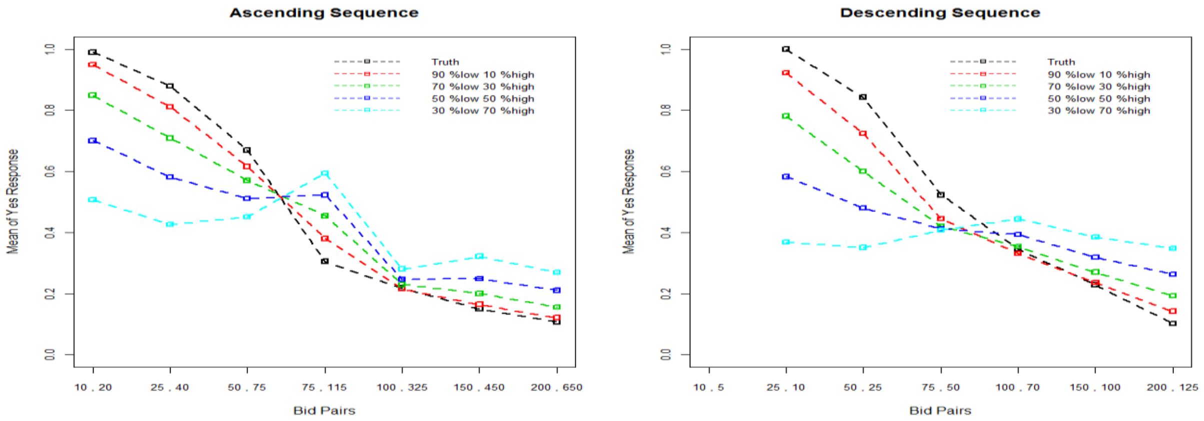


Figure 4: Partial Sample Anchoring

Figure 4a: Contemporaneous Anchoring: First Bid

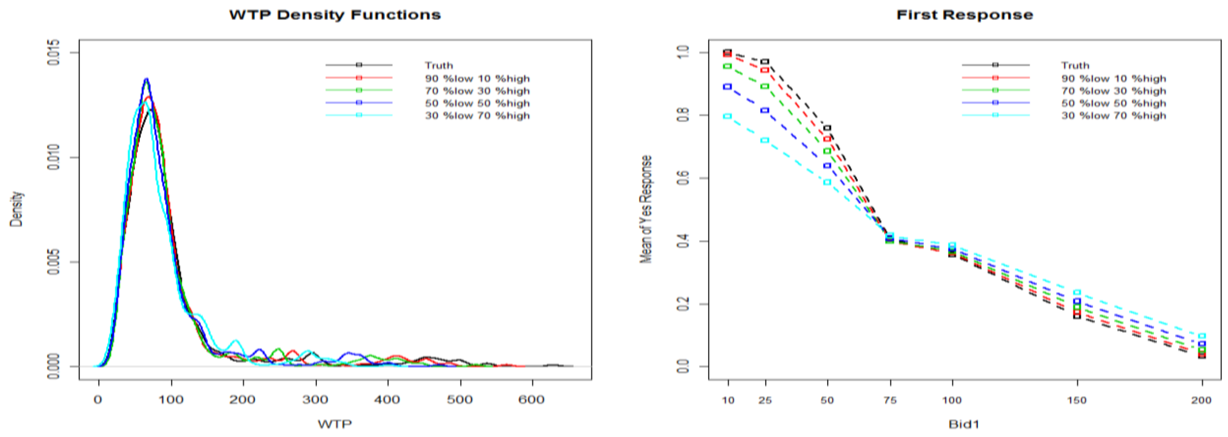


Figure 4b: Contemporaneous Anchoring: Second Bid

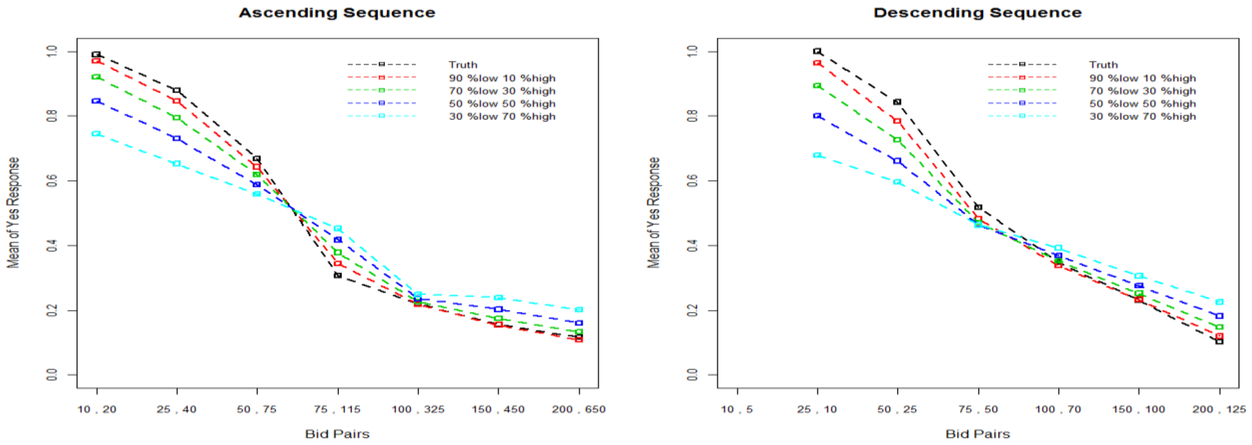


Figure 4c: Lagged Anchoring: First Bid

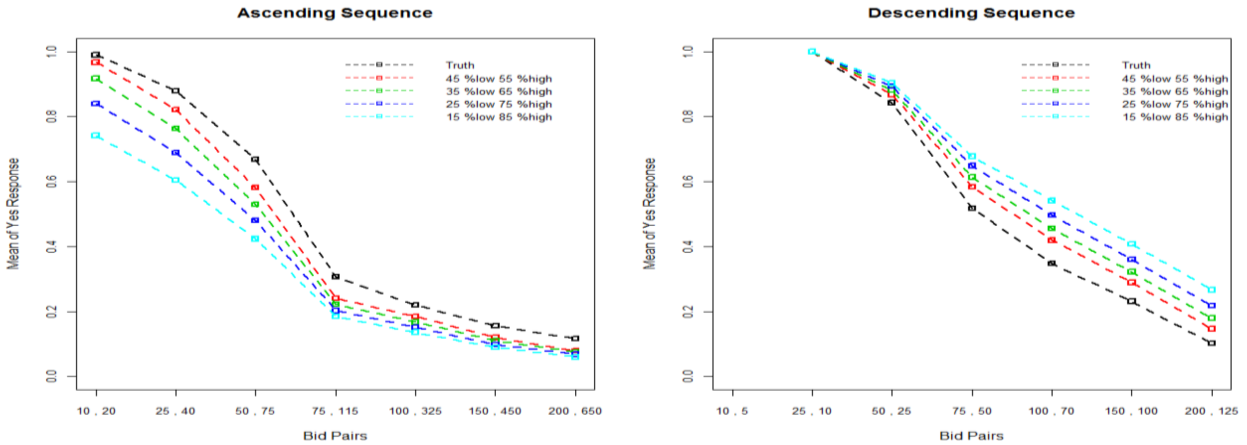


Figure 5: Full Sample Yea Saying

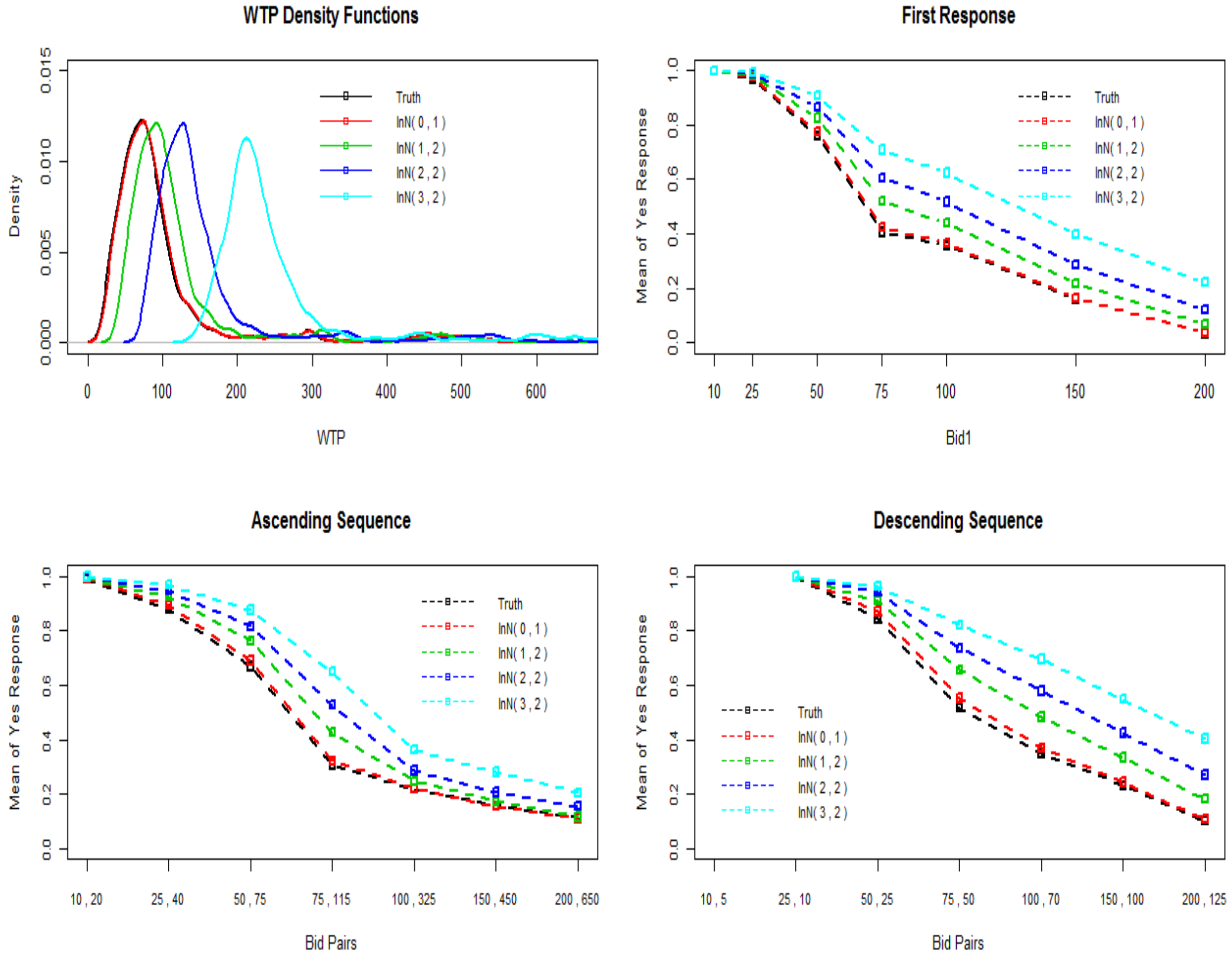


Figure 6: Partial Sample Yea Saying

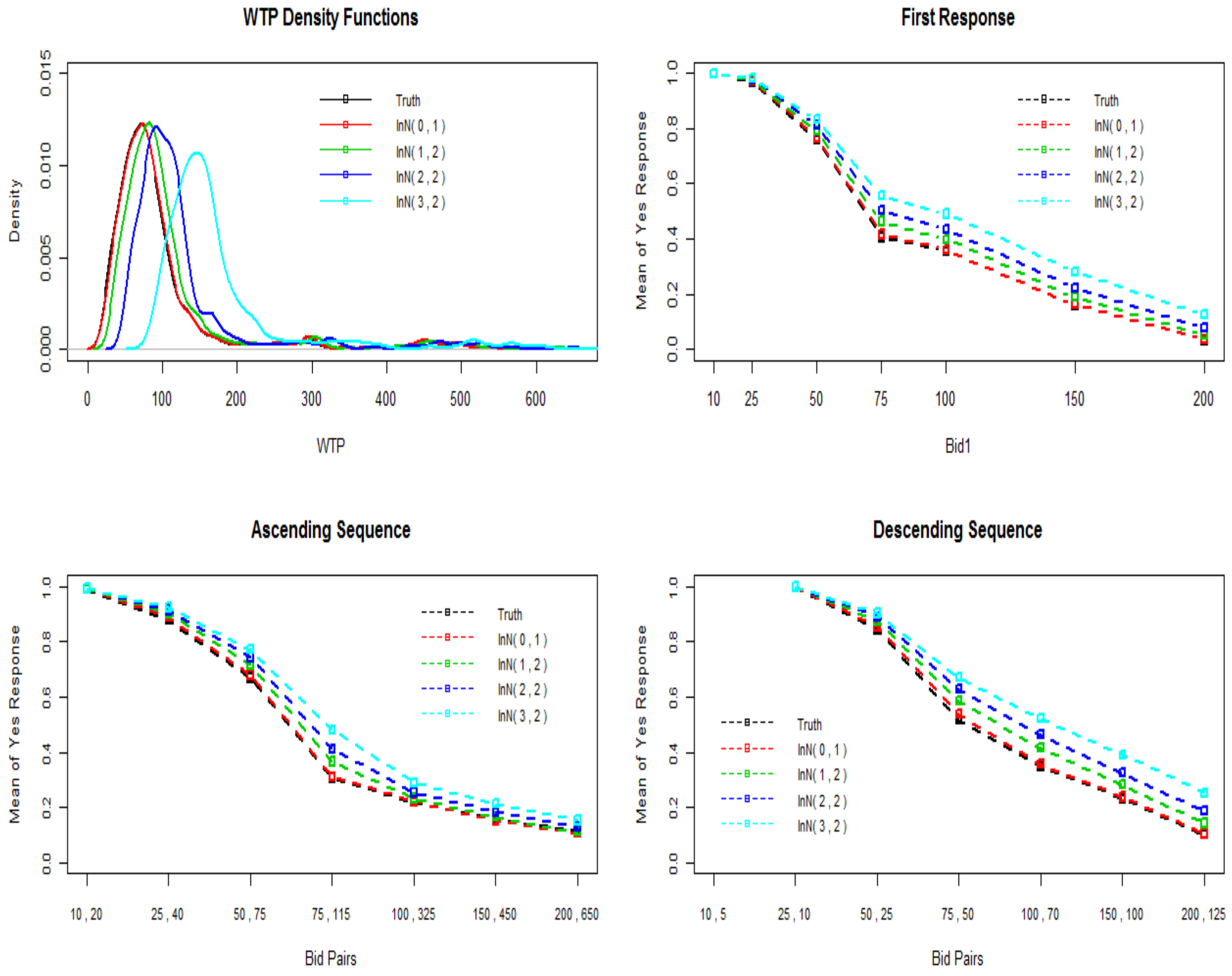


Figure 7: Other Elicitation Effects

Figure 7a: Full Sample Nay Saying

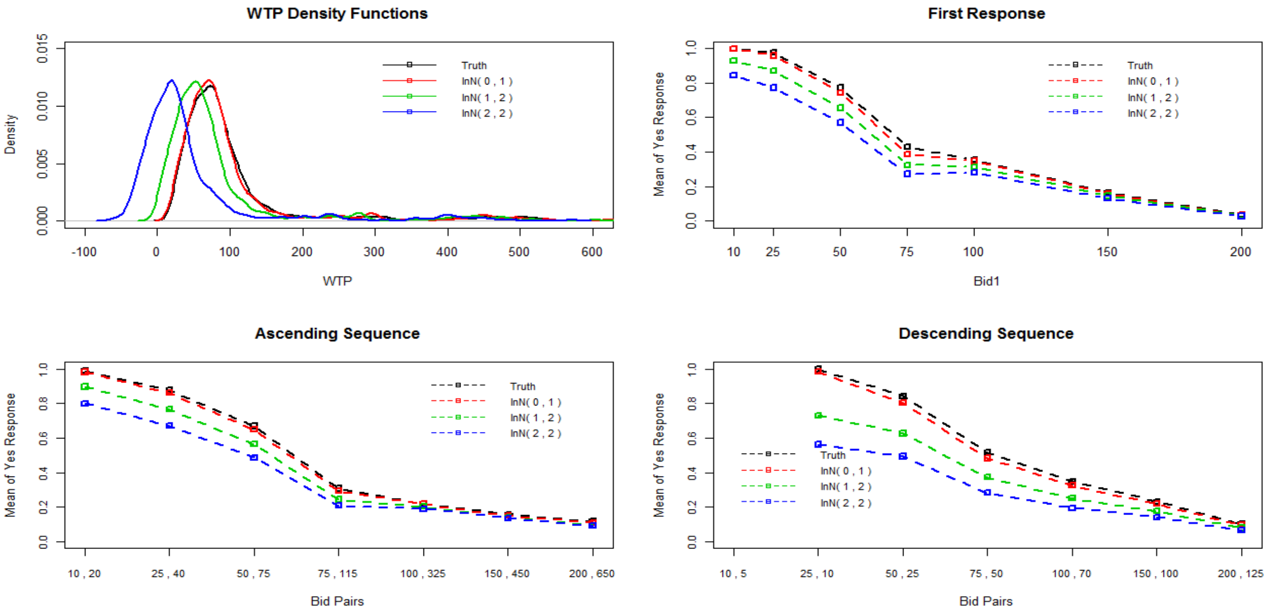


Figure 7b: Framing

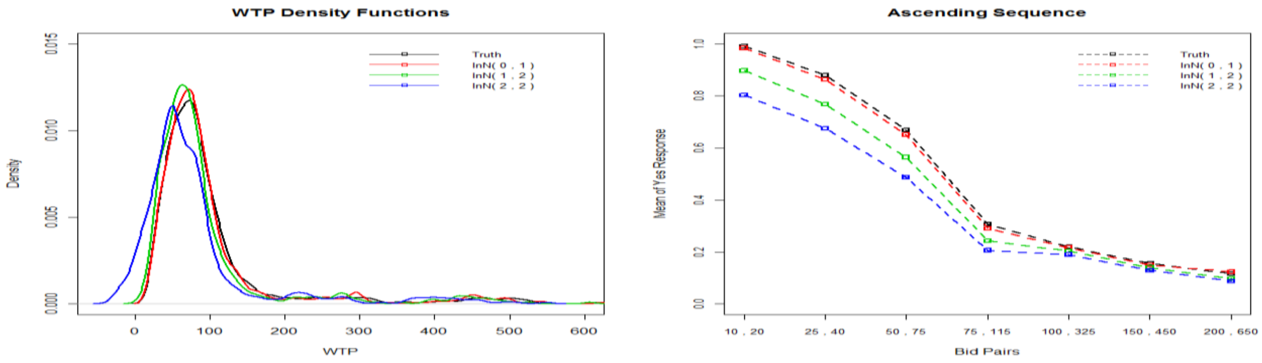


Figure 7c: Preference Uncertainty with Anchoring

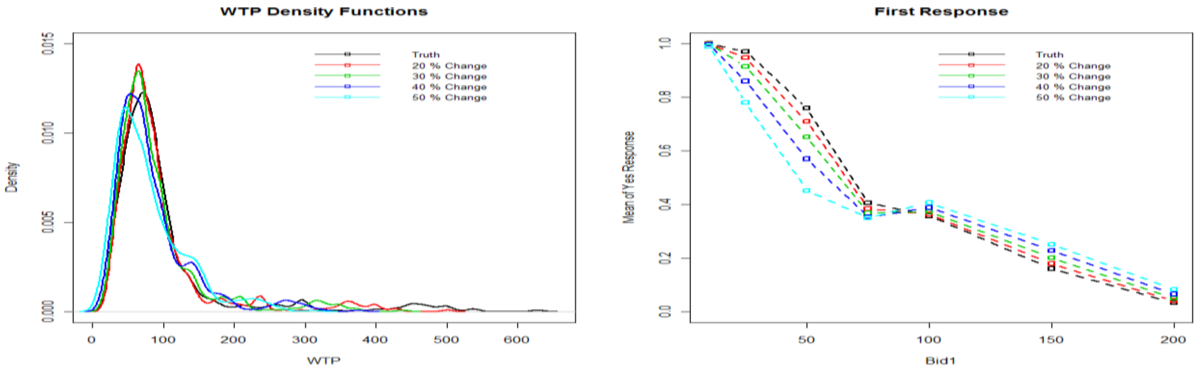


Figure 8: Robustness Checks w/ CES Utility Function

Figure 8a: WTP Density Functions

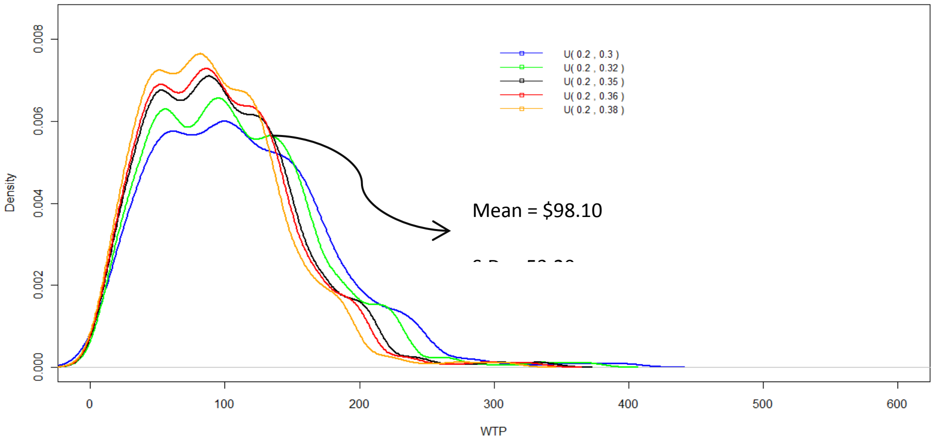


Figure 8b: First Bid Contemporaneous Anchoring

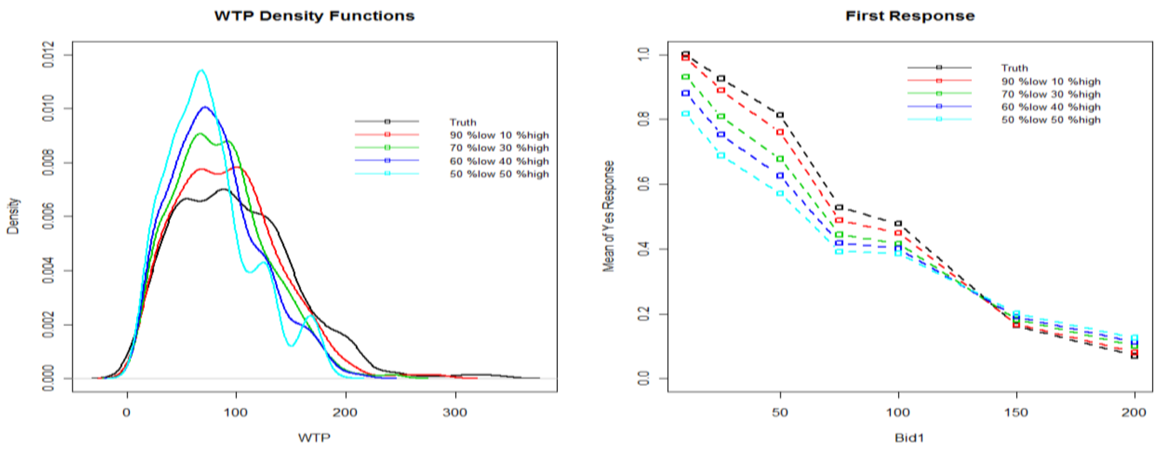
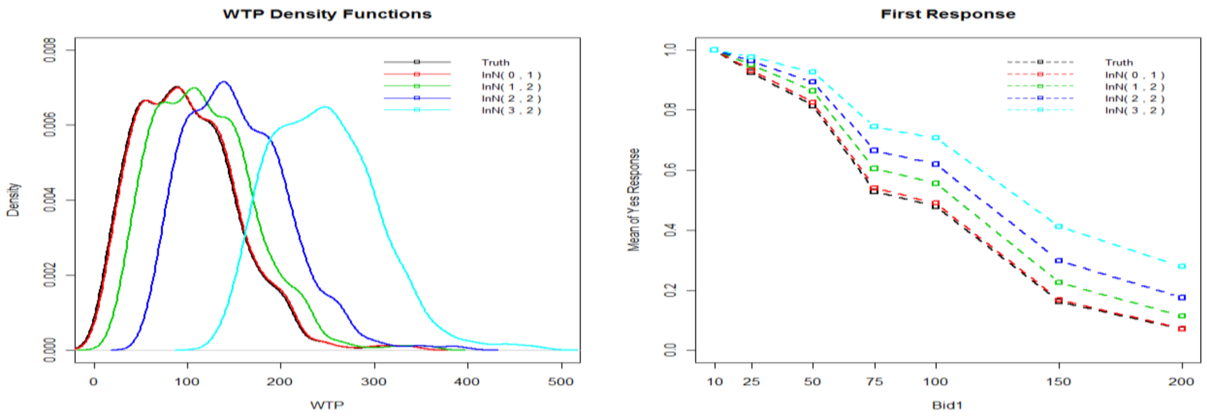


Figure 8c: Yea Saying



Chapter IV. Does Location Matter? The Effect of Institutional Versus Home Based Deliveries on Maternal Health

Abstract

In developing countries, home-based deliveries are carried out by traditional birth attendants who use outdated equipment and procedures. Mothers can suffer from physical and psychological consequences from these traditional delivery practices. These consequences can reduce a women's health and the treatment of these consequences can reduce household income, which is also known to adversely affect women's health. We examine the effect of delivery location (home versus health institution) on mothers' health (measured by body mass index). Using the program evaluation approach, we estimate average treatment effect on the treated for two Indian states, Bihar (the poorer state) and Gujarat (the richer state). Our propensity score matching results show that mothers in Bihar who use health institutions for child deliveries tend to be healthier than those who deliver at home. The richer state, Gujarat, where more advance home-delivery practices may be employed shows insignificant treatment effects. We recommend that the quality of home-based deliveries should be improved in poorer states like Bihar. Additionally, policies that increase accessibility and promote demand for/of institutional deliveries in poor states could be evaluated for their effectiveness and potential enhancement.

Key Words: place of childbirth, propensity score matching, program evaluation, average treatment effect on treated, maternal health

1. Introduction

Achieving universal access to reproductive health is one of the United Nations Development Programme's top eight Millennium Development Goals to be achieved by 2015.[24] A United Nations summit in 2010 found that *"(l)arge disparities still exist in providing pregnant women with antenatal care and skilled assistance during delivery ... this is especially true for regions where the number of skilled health workers remains low and maternal mortality high — in particular sub-Saharan Africa, Southern Asia and Oceania"*.[25] Insufficient access to health care services in these regions potentially explains why 90% of all global maternal mortality and morbidity take place in developing countries.[26]

Maternal health care includes services provided to mothers during pregnancy, childbirth and postpartum periods. In this paper, we focus on the location of childbirth (health institution versus home-based) as a proxy for health care services received during childbirth. In developing countries, traditional birth attendants perform home-based deliveries using traditional practices that are often harmful for mothers and children (Schairder et al., 1999; Goodburn et al., 2000). Not surprisingly, these home-based deliveries increase the risks of mortality and morbidity in mothers. Studies have examined the impact of location of childbirth on maternal mortality in developing countries (Barnett et al., 2008; Hogan et al., 2010; Vora et al., 2009). However, no study, to our knowledge has investigated the impact of delivery location on the health status of mothers and our research is aimed at filling this gap in the literature.

[24] http://www.undp.org/content/undp/en/home/mdgoverview.html, accessed March 2013.

[25] http://www.un.org/millenniumgoals/pdf/MDG_FS_5_EN_new.pdf, accessed March 2013.

[26] http://web.worldbank.org/WBSITE/EXTERNAL/TOPICS/EXTHEALTHNUTRITIONANDPOPULATION/EXTPRH/0,,contentMDK:20201062~menuPK:548481~pagePK:148956~piPK:216618~theSitePK:376855,00.html, accessed March 2013.

We use the body mass index (*BMI*) as an indicator of mother's health status as recommended by NIH.[27] Unskilled delivery assistance can have several physical (e.g. disabilities, damage in pelvic structure, severe acute maternal morbidity, anemia, sepsis, postpartum infections, hemorrhage etc.) and psychological (e.g. depression, stress etc.) consequences (Bang et al., 2004; Bentley and Griffiths, 2003; Filippi et al., 2006; Goodburn et al., 2000; Koblinsky et al., 2012). These consequences can potentially reduce mothers' *BMI*. Households incur financial losses in the process of treating the adverse physical and psychological impacts. Financial losses can then further increase the risk of weight loss in mothers since income is positively associated with the women's body weight in developing countries (McLaren, 2007).

While estimating the relationship between delivery location and mother's *BMI* is critical for all developing countries, we focus on India in this study for three reasons.[28] First, maternal health care is of a great concern as India is estimated to have acute levels of maternal morbidity and mortality as compared to other developing countries (Center for Reproductive Rights, 2008). Second, delivery care in India is highly inadequate and inequitable. According to India's national family health report, 60% of children are born at home. More than half of the children born at home are delivered by traditional birth attendants, relatives or other untrained persons (IIPS and Macro International, 2007). Third, close to 50% of Indian women are underweight (IIPS and Macro International, 2007). These are the women who face a high risk of maternal health related complications (Koblinsky, 1995; Sebire et al., 2001).

[27] The National Institute of Health and the National Heart, Lung and Blood Institute recommend using *BMI* as a predictor of health (NIH, 1998).

[28] India's National Family Health Surveys (NFHS) provide information on a variety of family health and welfare indicators that allow us to empirically estimate the effects associated with institutional deliveries. These data are available upon request from The Demographic and Health Surveys (http://www.measuredhs.com/).

We estimate health effects of institutional deliveries for two socio–economically and spatially diverse Indian states, Bihar and Gujarat. Bihar, an eastern state, is one of the poorest states in India. Only 9% of the households belong to the highest wealth quintile and only one in five births take place at a health facility (IIPS and Macro International, 2008a). In contrast, Gujarat, a western state, is one of the fastest growing states in the country. Thirty two percent of households in Gujarat belong to the highest wealth quintile, and more than half of children in Gujarat are delivered at health institutions (IIPS and Macro International, 2008b). We estimate separate treatment effects for each state.

The program evaluation approach is used to estimate the treatment effects associated with the location of childbirth. We match the treated group (mothers who use health institution for child births) and control group (mothers who deliver child at home) using propensity scores that are defined as the conditional probability of receiving a treatment, ceteris paribus (Dehejia and Wahba, 2002; Rosembaum and Rubin, 1983). Different matching algorithms (i.e. Caliper and Nearest Neighbor) are used to estimate the treatment effects. This allows us to investigate the robustness of treatment effect estimates.

We find that mothers from educated, wealthy and urban households are more likely to have institutional deliveries. We find a positive and significant treatment effect associated with institutional deliveries for mothers in Bihar (the poor state), but no significant treatment effect is found for Gujarat (the richer state). This difference in estimated treatment effects is potentially an outcome of differences in the socio-economic status of households and the quality of home-based deliveries across states. Our findings suggest that the quality of home-based deliveries should be improved in poorer states like Bihar and that services provided in richer states might provide models for these improvements. Additionally, the proportion of institutional deliveries should be increased by targeting the demand as well as the supply of maternal health care services.

2. Program Evaluation of Delivery Location

To our knowledge, no study has empirically estimated the effect of delivery location on mothers' post-delivery, long-term *BMI*. To examine this relationship, we use the treatment evaluation approach that is commonly used for impact and policy assessment using observational and experimental data (Heinrich, Maffioli, Vazquez, 2010). [29] Estimation of treatment effects using observational data may be limited due to selection bias and endogeneity. These unintended consequences arise because the treatment (e.g. delivery location) is not randomly assigned. The treatment evaluation approach overcomes the selection bias by creating a counterfactual control group which is similar to the treated group so that difference in the outcomes of these two groups is explained solely by the treatment assignment.

We create the control group using propensity score that is a measure of the probability of choosing to receive a treatment. We draw insights from existing literature to model the choice of delivery location in India. Socio–economic household indicators like wealth and urban residence; and mother's education and age are found to be positively associated with the use of institutional maternal health services (Griffiths and Stephenson, 2001; Thind et al., 2008; Naveentham and Dharmalingam, 2002; Padmadas et al., 2000). The birth order of child is an important determinant of use of skilled delivery assistance in India (Thind et al., 2008; Naveentham and Dharmalingam, 2002). Additionally, religion of mothers can explain the choice of delivery

[29] For a detailed overview of program evaluation approach and propensity score matching, readers are directed to Caliendo and Kopeinig (2008), Heinrich, Maffioli, Vazquez (2010), and Chapter 25 in Cameron and Trivedi (2005).

location (Naveentham and Dharmalingam, 2002). We use these variables based on their significance in prior research.

The binary treatment variable for our analysis is:

$$D_i = \begin{cases} 1\, child\ delivered\ at\ a\ health\ facility \\ 0\, child\ delivered\ at\ home \end{cases}, \qquad i = 1, .., N, \qquad (1)$$

where *N* is the number of mothers. Suppose that the potential health outcome (measured by body mass index) of this treatment variable for mothers is shown by the vector y_{iD_i}, where D_i takes the value "1" for treated mothers and "0" for non-treated mothers. Ideally, we would like to estimate the effect of treatment on the health outcome of each participant i.e. $\pi_i = y_{i1} - y_{i0}\ \forall\ i$. However, π_i cannot be identified as y_{i1} and y_{i0} are not observed for the same participant. To counter this identification problem, we estimate the sample analogues of individual treatment effects.

The sample analogues are the average treatment effect on the treated (ATT) and average treatment effect (ATE) that are defined as:

$$ATT = E(y_{i1} - y_{i0}|X_i, D_i = 1) = E(y_{i1}|X_i, D_i = 1) - E(y_{i0}|X_i, D_i = 1), \qquad (2)$$

$$ATE = E(y_{i1} - y_{i0}|X_i, D_i) = E(y_{i1}|X_i, D_i = 1) - E(y_{i0}|X_i, D_i = 0), \qquad (3)$$

where X_i includes the demographics characteristics of the mothers such as those described above. ATT defined in equation (2) measures the average gain in health outcomes for mothers who can use health institutions for deliveries whereas ATE in equation (3) measures the average gain in *BMI* for the entire sample.

ATE is relevant measure only if the assignment is universally affordable and accessible in which case a randomly selected population can depict the average impact of the treatment on health outcomes (Chapter 25, Cameron and Trivedi, 2005). In our analysis we focus on the estimation of ATT since maternal health care is not universally accessible and affordable in India. The average treatment effect on the treated given in equation (2) can be written as:

$$ATT = ATE - [E(y_{i0}|X_i, D_i = 1) - E(y_{i0}|X_i, D_i = 0)] = ATE - SB, \qquad (4)$$

where SB is the selection bias that originates because of the inability to control for the differences in treated and control group. The objective of empirical analysis is to minimize the selection bias so that the ATT can be approximated using ATE. In experimental trails, treatment is random and there is a well-defined control group. The outcomes are only affected by the assignment of the treatment in which case SB in equation (4) equals zero.

In observational data, the differences in health outcomes of the treated and non-treated participants are affected by the treatment assignment and the characteristics (observables and unobservable) of the participants implying there could be potential selection bias. Also, there is a fundamental identification problem because of the inability to observe the outcome for the treated in absence of the treatment shown by $E(y_{i0}|X_i, D_i = 1)$ in equation (2). To counter these two issues, we resort to matching that uses the pool of non-treated individuals to create a control group, which is characteristically similar to the treated group. Similarity can be addressed by matching the observed characteristics of individuals. However, the choice of delivery location can be explained by several factors in which case matching based on characteristics may not be feasible because a large number of covariates can slow down the estimation procedure and give rise to the curse of dimensionality. Therefore, we employ propensity score matching that uses the probability of choosing to receive a treatment to match the treated and non-treated groups.

2.1. Propensity Score Matching

The intuition behind matching is to find similar treated and non-treated (control) groups. Rosenbaum and Rubin (1983) introduced the concept of propensity score matching. A propensity score is a scalar that measures the conditional probability of receiving a treatment given the matrix of characteristics, X_i. In mathematical notations, it is defined as:

$$p(x) = Prob[D_i = 1|X_i]. \tag{5}$$

We use the logistic distribution $\left(Prob[D_i = 1|X_i] = \frac{1}{1+e^{-X_i'\beta}}\right)$ to estimate propensity score for each observation. For matching, the predicted value of equation (5) i.e. $\hat{p}(x) = F\left(X_i'\hat{\beta}\right)$ or predicted odds ratio $\left(\frac{\hat{p}(x)}{1-\hat{p}(x)}\right)$ can be used as weights to create the control group of untreated mothers.

2.2. Matching Algorithms

A number of matching algorithms have been proposed in the literature. In this paper we use the Nearest Neighbor and Caliper matching tools.[30] For a detailed explanation of these matching techniques, readers are directed to Rosenbaum and Rubin (1983), Heckman et al. (1998), Dehejia and Wahba (2002), Smith and Todd (2005) and Caliendo and Kopeining (2008). Below, we briefly explain the matching algorithms.

[30] We use the “psmatch2” command in STATA 12 to estimate ATT and ATE via propensity score matching. Refer Leuven and Sianesi (2003) for more details.

Nearest neighbor matching uses the closest propensity score to match a treated participant with a non-treated participant to create a control group. Suppose that $p_i, i \in T$ is the estimated propensity score of the i^{th} treated individual. The closest non-treated neighbor is such that the difference $L_i = \|p_i - p_j\|$ is minimized with respect to j, where $j \in NT$. With one-to-one matching only the first nearest neighbor is considered. With multiple neighbors, more than one neighbor is used to create the control group for i^{th} treated individual. Matching with multiple neighbors is known to reduce the variance at the cost of increasing bias. Additionally, matching can be performed with and without replacing the non – treated observations. Matching with replacement is preferred as it reduces the estimation bias. We use one-to-one and multiple matching with replacement to estimate the treatment effects.

A limitation of the Nearest Neighbor procedure is that it can potentially construct a control group even if the propensity scores differ significantly. To resolve this, we estimate ATT by Caliper matching that limits matching to a maximum tolerance. The closest non-treated neighbor is such that the difference $L_i = \|p_i - p_j\| < r$ is minimized, where *r* denotes the radius or caliper for matching. A priori it is difficult to know which values of *r* should be used. In this analysis different values are assigned to r for testing the robustness of ATT estimates. Huber, Lechner and Wunsch (2013) compare several matching techniques and find that Caliper matching performs best overall.

With Nearest Neighbor and Caliper matching algorithms, ATT can be estimated in the following fashion. Suppose that *C(i)* is the control group created for the i^{th} treated participant and N_i^C be the number of observations in *C(i)*. With matching, ATT given in equation (2) is estimated as:

$$ATT = \frac{1}{N^T}\sum_{i\in T}\left[y_{i1} - \sum_{j\in C(i)} w_{ij}y_{i0}\right] = \frac{1}{N^T}\left[\sum_{i\in T} y_{i1} - \sum_{i\in T}\sum_{j\in C(i)} w_{ij}y_{i0}\right], \quad (6)$$

where N^T is the number of treated observations; and $w_{ij} = \frac{1}{N_i^C}$ if $j \in C(i)$ and $w_{ij} = 0$, otherwise. The variance of ATT is calculated as:

$$\sigma^2 = \frac{1}{N^T}Var(y_{i1}) + \frac{1}{(N^T)^2}\sum_{j\in C(i)} w_{ij}^2 Var(y_{i0}). \quad (7)$$

We can test the quality of matching via propensity scores using two procedures. First, the estimated propensity scores should meet the common support criterion, which requires an overlap between the propensity scores of treated and the matched control groups. This implies that mothers with same characteristics should have positive probabilities of receiving and not receiving the treatment, and the estimated probabilities should be distributed between 0 and 1. Simple inspection of histograms of propensity score for the treated and control groups can help to investigate whether the overlap condition is met.

Second, we examine whether the balancing condition is met by propensity score matching. The balancing condition requires that the covariates across treated and control groups are similar after matching. We estimate the percentage bias for each covariate before and after matching as:

$$\%Bias = \frac{\bar{X}_T - \bar{X}_{NT}}{\sqrt{\frac{\sigma_T^2 + \sigma_{NT}^2}{2}}} \times 100, \quad (8)$$

where T denotes the treated group and NT stands for the non–treated group. $\bar{X}$ and σ^2 represent the mean and variance of these groups, respectively. A t–test for difference in means is used to test whether $\%Bias$ in covariates across two groups is significant.[31]

3. Data Description

The Center for Reproductive Rights (2008) states that *"(m)ore women die due to pregnancy-related causes in India than anywhere else in the world"*. Additionally, the report found that *"the incidence of maternal morbidity is very high, making it an equally pressing concern"*. A potential reason for high maternal mortality and morbidity rates is that the access to health care services is highly inequitable and lower socio – economic groups face the greatest difficulties in meeting their health needs (Balarajan, Selvaraj, Subramanian, 2011; Subramanian and Smith, 2006). For example, half of Indian mothers do not receive three antenatal visits during pregnancy and only 53% of births are assisted by trained skilled health workers.[32]

Delivery location is an important determinant of overall maternal health. Skilled delivery assistance during delivery can potentially affect maternal mortality and morbidity. In India, a high incidence of maternal mortality has been attributed to home-based deliveries (Vora et al., 2009; Barnett et al., 2008; Basu and Stephenson, 2005; Gupta, 1990). For example, Vora et al. (2009) find that "(*d)eaths due to sepsis and obstructed labor may be attributed to the high proportion of deliveries at home"*. Similarly, Barnett et al. (2008) find that in Jharkhand and Orissa, 87% of mothers who died because of maternal health conditions, delivered at home.

[31] Estimation of treatment effect via propensity score matching also requires that the conditional independence assumption is met. This condition requires that the potential outcomes are independent of the treatment, given the propensity scores i.e. $y_{i0}, y_{i1} \perp D|p(x), \forall x$. However, this assumption is not directly testable as the true distribution of y_0 given that they received the treatment is unknown (Imbens and Rubin, 2010).

[32] http://www.unicef.org/india/health.html

Data to address institutional versus at-home maternal care are from Bihar and Gujarat states, and are drawn from the second and third rounds of National Family Health Survey (NFHS) that were collected in 1998-99 and 2005-06, respectively.[33] The International Institute for Population Sciences, Mumbai in coordination with several other organizations conducted these in person surveys. The objective of NFHS was to collect data on health and family welfare from representative households in all the 26 Indian states using stratified, random techniques.[34] The NFHS data were collected at three levels: village, household and individual women. In our study, we use data collected on maternal and household characteristics via the women questionnaire.[35]

The 1998-99 NFHS collected information from over 7,000 and 3,800 women in Bihar and Gujarat, respectively. The 2005-06 NFHS collected information on over 3,800 and 3,700 women in Bihar and Gujarat, respectively. We exclude data on women who did not deliver any children in five years preceding the interviews, women who were pregnant or had tuberculosis at the time of the interviews, and women with missing data on any of the variables used in our analysis. In addition, we excluded data on women (less than 1% of observations) who belonged to Christianity, Sikhism and Buddhism in Gujarat. This was done because there were no followers of these religions in the final Bihar data. With these exclusions, the 1999-98 NFHS provides information on 2,176 and 937 mothers in Bihar and Gujarat, respectively and the 2005-06 NFHS results in 1,053 mothers in Bihar and 868 mothers in Gujarat.

[33] We do not include data from the first round of NFHS in our analysis because of the unavailability of data on mother's weight and height.

[34] The sampling techniques for the two rounds of NFHS are available at http://www.rchiips.org/nfhs/NFHS-3%20Data/VOL-2/Appendix%20C.pdf and http://www.rchiips.org/NFHS/data/bh/bhchap1.pdf.

[35] The women survey instrument for NFHS 1998 – 1999 can found at http://www.rchiips.org/NFHS/data/bh/bhwomqre.pdf. The survey instrument for NFHS 2005 – 2006 is available at http://www.rchiips.org/nfhs/NFHS-3%20Data/VOL-2/Woman%20Questionnaire.pdf.

Summary statistics of variables used in our analysis are given in Table 1. The first variable, DELIVERY is our treatment variable that indicates the location of delivery (home vs. health care institutions) of last child born in five years preceding the interviews. From the 1998-99 to 2005-06 the proportion of mothers using health institutions for child deliveries increased by 12% and 7% in Bihar and Gujarat, respectively. In 2005-06 only 26% of the mothers in Bihar used health institutions for deliveries, which is substantially lower than that of Gujarat (57%).

The health outcome is BMISTATUS, which is defined as:

$$BMISTATUS = \begin{cases} 0 \; if \; BMI < 18.50 \\ 1 \; if \; BMI \geq 18.50 \end{cases}, \tag{9}$$

where *BMI* is the body mass index calculated as $\frac{\text{Weight in Kilograms}}{\text{Height in Meters}^2}$. It is common to classify BMI using three categories – underweight (*BMI < 18.50*), normal weight (*18.50 ≤ BMI < 25*) and overweight (*BMI > 25*). We combine the normal weight and overweight categories because only small percentages of mothers are overweight, 3% in Bihar and 7-10% in Gujarat, which precludes investigating separate effects for this group in each state.[36]

The percentage of mothers with *BMI* greater than 18.50 decreased by 5% between the second and third survey phases in Bihar (Table 1). This was an unexpected result because the use of maternal health facilities nearly doubled over this period of time in Bihar. This outcome is explained by the distribution of *BMI* in 2005-06 being more dispersed and skewed than in 1998-99 (Figure 1). The average *BMI* between the two survey phases is not significantly different whereas the median *BMI* is different at the 10% level of significance. For Gujarat this dispersion

[36] Additionally, in 2005-06 less than 0.05% and 2% of mothers are obese (BMI≥30) in Bihar and Gujarat, respectively. If we drop observations of the overweight category (BMI≥25), the estimated treatment effects do not change.

increase and skewness is also present (Figure 2), but the percentage of mothers with *BMI* greater than 18.50 did increase slightly from 1998-99 to 2005-06.

For propensity score estimation, we include variables that potentially influence delivery location choices. Variables from three categories are used in the logit modeling: mother's, household's and child's characteristics. Mother's characteristics include variables such as AGE, SINGLE and WORKING. There are several reasons to expect greater proportion of institutional deliveries for older mothers. First, they can potentially face greater health risks by delivering children at home. Second, older mothers are more likely to have more power in household decision making as compared to younger mothers. Third, older women likely have more child-birth experience. In both states, the average age of mothers increased across the two survey phases.

Mothers who are single can favor home-based deliveries because of accessibility, convenience and lower expense. In contrast, working mothers may prefer institutional deliveries because they have income and may not want risk losing their employment time from birth complications. However, most working mothers in Bihar and Gujarat are employed in the agricultural and manual labor sectors that provide low wages and do not provide family-leave time. For these mothers, institutional deliveries come at an opportunity cost of forgoing labor incomes. In both survey phases, the proportion of single mothers was greater in Bihar whereas greater percentages of mothers were working in Gujarat.

We include the sex of the household head as an explanatory variable as it can shed light on the power structure of the household. Female-headed households may provide impetus for mothers to use institution based maternal health services. The majority of households in Bihar and Gujarat had male heads.

Household education may positively affect the choice of institutional deliveries for childbirths. We use minimum education of the husband and wife to indicate household education.[37] On average, we find that households in Bihar tended to be less educated than those in Gujarat.

The wealth of households is an important determinant of accessibility. In India, the lower socio-economic groups face the greatest difficulties in meeting their health needs (Balarajan, Selvaraj, Subramanian, 2011; Subramanian and Smith, 2006). The variable, WEALTHINDEX is a consolidated index of over 30 household assets (e.g. ownership of items, source of water, access to electricity etc.).[38] This index divides individuals into five equal wealth quintiles (i.e. POOREST, POORER, MIDDLE, RICHER, RICHEST).

Unfortunately, the wealth index is not available for 1998-99 NFHS. So we use two other variables to indicate wealth of households. The variables, TOILET and ELECTRICITY, show the proportion of households with access to toilets and electricity connections, and may also indicate wealthier households. In addition, these variables can potentially result in safer home-based deliveries. These two variables along with WEALTHINDEX variables show that the economic status of households in Gujarat was better than those in Bihar.

We also incorporate if mothers live in an urban area (RESIDENCE) and RELIGION of households in propensity score estimation. We might expect mothers that live in urban areas to

[37] We also investigated specifications using maximum and average household education, the results are qualitatively similar.
[38] Household Characteristics (Page 2), National Family Health Survey 2005 – 2006, India: Key Findings (2007).

be more likely to use institutional health-care facilities. Existing evidence shows that in some states Muslim women are potentially less likely to have institutional deliveries as compared to Hindu mothers. For example, Naveentham and Dharmalingam (2002) find that "*Muslim women in Kerala were 70% less likely to deliver a baby in a heath care institution than Hindu women*".

Lastly, the birth order of the child can affect the choice of location of delivery since first time mothers are inexperienced and most likely to seek professional assistance for childbirths (Naveentham and Dharmalingam, 2002; Thind et al., 2008).

4. Results

We start with a discussion of the propensity score results and then analyze our matching results.

4.1. Propensity Score Estimation

Three patterns are seen in our logistic regression results. First, three variables significantly affect the dependent variable, DELIVERY across states for both survey phases. Households with the highest minimum education are more likely to use health care institutions for child deliveries as compared to those who have no education. Urban households have a greater probability of choosing health care institutions as compared to rural households and first time mothers are most likely to go for institutional deliveries. If the last child has a higher birth order, it will decrease the likelihood of using health care services for delivery.

Second, five variables are significant for at least three of the four state/survey version equations. AGE, SECONDARY EDUCATION and TOILET increase the probability of institutional deliveries, and WORKING decreases this probability.

Third, three variables are significant for only one state/survey version. RELIGION is significant for Bihar only. We find that Hindu mothers are more likely to go to health facilities for childbirths as compared to Muslim women. PRIMARY EDUCATION and ELECTRICITY are only significant in the equations for the 1998-99 version of the survey in both states. Overall, our results are in line with previous findings on the determinants of location of delivery in India (Naveentham and Dharmalingam, 2002; Padmadas et al., 2000; Thind et al., 2008).

We test whether the coefficients of the logistic regression are equal across states using the procedures outlined in Swait and Louviere (1993). For the 1998-99 and 2005-06 NFHSs the log likelihood ratio tests are significant at 5% and 10% levels, respectively. These tests indicate that significant differences exist between the coefficients and the scales, and consequently, the choice to use institutional delivery across states.[39]

4.2. Matching Results

Using the Nearest Neighbor and the Caliper matching algorithms mothers who use a health care facility for childbirths are matched with those who deliver at home.[40] For the Nearest Neighbor algorithm, we use one-to- one matching (*N=1*) and matching with multiple neighbors (*N=4*) with replacement (see Table 3). For comparison purposes, we use *r=0.01* and *r=0.05* in Caliper matching (see Table 3). We test the robustness of ATT results using multiple values of *r* in Table 4.

Before examining the treatment effect estimates, we discuss whether the conditions required for propensity score matching are satisfied. The first requirement is that there should be

[39] We only test for differences across state for each survey phase. The logistic regressions across survey years do not have the same specification for each, which precludes comparing the regression results.
[40] We used *"psmatch2"* command in STATA to estimate the treatment effects via propensity score matching.

an overlap between the estimated propensity scores of the treated and non – treated groups. For brevity, we present the overlap of propensity scores using the Nearest Neighbor ($N=1$) and Caliper Matching ($Tol = 0.01$) algorithms (Figure 3 and Figure 4). Other variants of matching algorithms provide similar insights.

Figure 3 plots the frequency histograms of propensity scores for the treated and non-treated mothers in Bihar and Gujarat using Nearest Neighbor matching. The distributions overlap, but as expected the distribution of the non- treated group is highly skewed toward lower probabilities of using institutional delivery. We find that the satisfying the overlapping condition is robust to trimming of potential extreme observations. [41]

We find that Caliper matching generates a few off support (no overlap) observations. In Figure 4(b) for Bihar, over 8% of the treated mothers do not find an overlap with the non- treated group of mothers. Similarly, around 6% of the non- treated mothers do not generate an overlap with the treated mothers. Similarly, for Gujarat about 6% of treated observations do not find an overlap with the non- treated group. The off support observations are excluded from estimation but that has no effect on the estimated ATT.

Second, we test for the balancing condition. For brevity, we present graphs on covariate balance for Bihar and Gujarat using NFHS 2005-06 data for the Nearest Neighbor ($N=1$) matching algorithm. Consider Figure 5, the x-axis contains the percentage bias given in equation (8) for covariates presented on the y-axis. For Bihar, $\%Bias$ is significant for all variables before matching.[42] The mean $\%Bias$ of all covariates is about 65% and, the mean bias reduces to 6%

[41] To further test the sensitivity of our treatment effect estimates, we trim observations that have a propensity score of 0.6 or higher. This accounts for about 10% of the total observations. Trimming of the observations does not change our findings.

[42] We tested the significance of %Bias using t – test.

after matching. For Gujarat, the mean $\%Bias$ of all covariates is 48% before matching and after matching the mean $\%Bias$ reduces to 4%. Similar insights emerge for Bihar and Gujarat using Caliper matching.

The treatment effect results are summarized in Table 3. We find that ATT is positive and significant for Bihar using NFHS 2005-06 data only. All other ATTs presented in Table 3 are insignificant and the ATT estimates are robust to the four matching algorithms used. Since we cannot a priori select the optimal value of r, we use multiple values to test the robustness of ATT for mothers in Bihar in 2005-06 (see Table 4). Since propensity scores are skewed, with very low tolerance (*r=0.0001*) only 19% of observations in the treated group are used for ATT estimation. Thus, the resulting ATT is insignificant. As expected, reducing the tolerance increases the number of treated observations for estimating ATT. With *r=0.005* over 80% of the treated observations are included for calculation and resulting ATT is positive and significant.

4.3. Matching Discussion

The treatment effect estimates in Table 3 bring out some important insights. First, consider Bihar. For 1998-99 NFHS, only 14% of mothers in Bihar used health centers for child deliveries. The distribution of propensity scores for a majority of mothers is highly skewed (toward lower scores) which limits the variation in the matched control group. Thus, it's not clear whether the insignificance of ATT is driven by the inability to create a well-defined control group or the actual lack of an effect.

In 2005-06, the ATT is positive and significant for Bihar. Since we are comparing the outcome of underweight (base group) and normal mothers, a positive ATT implies that institutional deliveries tend to have a positive effect on the health of mothers, which is measured by the body mass index. Conversely, there is an average loss in *BMI* associated with home-based deliveries for underweight mothers in Bihar. In 2005-06, the mean of BMISTATUS of the non-treated group is 0.49 which is less than that of the treated group (0.72).

We find insignificant treatment effects for Gujarat for both survey rounds. This result is not surprising. First, empirical evidence indicates that the quality of home-based deliveries is superior in Gujarat than in Bihar. For example, 61% of babies were immediately wiped and wrapped after birth at home, 22% of mothers received trained assistance, 32% used disposable delivery kits for home-based deliveries and 29% of home-based births were followed by a postnatal checkup. In contrast, 38% of babies were immediately wiped and wrapped after birth at home, 12% of mothers received trained assistance, disposable delivery kits were used in 2% of home-based deliveries and only 6% of mothers who delivered at home received antenatal care in Bihar (IIPS and Macro International 2008a, 2008b).

Second, mothers in Gujarat who deliver at home are more likely to receive health and nutrition based education during the periods of pregnancy and childbirths. For example, 14% of mother who were covered by an Anganwadi (childcare center) received health and nutrition based education during pregnancy in Gujarat whereas only 0.2% of mothers who are covered by an Anganwadi received health and nutrition based education during pregnancy in Bihar (IIPS and Macro International 2008a, 2008b). Third, Gujarat households have more wealth and are much more likely to have toilets, and electricity (Table 1). Thus, adverse post-delivery maternal health outcomes as indicated by weight loss from home-based deliveries will be less likely in Gujarat.

5. Conclusions and Policy Implications

With over 40% of underweight women and only 26% institutional child deliveries, our results suggest there is a need to reevaluate maternal health care services in Bihar and other similar lower-income status states in India and elsewhere. To address this issue, we suggest two policy targets. First, there is a need to improve the quality of home-based deliveries in poorer states like Bihar. For example, mothers could be provided with skilled assistance during childbirth, disposable delivery kits or other proven services that influence maternal health to reduce post-delivery health risks and potential financial losses.

Promotion of policies that increase accessibility and promote demand for/of institutional deliveries in poor states could be evaluated for their effectiveness and potential enhancement. The propensity score results indicate that households that are educated, wealthy and reside in urban areas are more likely to go for institutional deliveries. This suggests that programs to enhance the overall economic condition of low-income households will have a spin-off effect of enhancing material postpartum health if institutional maternal services are available and financially accessible.

References

Bang R.A., A. T. Bang, M. H. Reddy, M. D. Deshmukh, S. B. Baitule, V. Filippi. 2004. Maternal Morbidity during Labor and the Puerperium in Rural Homes and the Need for Medical Attention: A Prospective Observational Study in Gadchiroli, India. International Journal of Obstetrics & Gynecology, 111(3): 231–238.

Barnett S., N. Nair, P. Tripathy, J. Borghi, S. Rath, A. Costello. 2008. A Prospective Key Informant Surveillance System to Measure Maternal Mortality – Findings from Indigenous Populations in Jharkhand and Orissa, India. BMC Pregnancy and Childbirth, 8:6.

Balarajan Y., S. Selvaraj, S.V. Subramanian. 2011. Health care and equity in India. The Lancet, 377 (9764): 505 – 515.

Basu A. M., R. Stephenson. 2005. Low Levels of Maternal Education and The Proximate Determinants of Childhood Mortality: A Little Learning is Not a Dangerous Thing. Social Science & Medicine 60 (9): 2011 – 2023.

Bentley, M.E., P.L. Griffiths. 2003. The Burden of Anemia among Women in India. European Journal of Clinical Nutrition, 57: 52–60.

Bloom, S.S., T. Lippeveld, D. Wypij. 1999. Does Antenatal Care Make a Difference to Safe Delivery? A Study in Uttar Pradesh, India. Health Policy and Planning, 14(1): 38-48.

Cameron A. C., P. K. Trivedi. Microeconometrics – Methods and Applications. 2009. Cambridge University Press, New York.

Caliendo M., S. Kopeinig. 2008. Some Practical Guidance for the Implementation of Propensity Score Matching. Journal of Economic Surveys, 22(1): 31-72.

Center for Reproductive Rights. 2008. Maternal Mortality in India – Using International and Constitutional Law to Promote Accountability and Change. http://www.unfpa.org/sowmy/resources/docs/library/R414_CenterRepRights_2008_INDIA_Maternal_Mortality_in_India_Center_for_Huiman_Rights.pdf, accessed March 2013.

Dehejia, R., Wahba, S., 2002. Propensity Score Matching Methods for Non-Experimental Causal Studies. Review of Economics and Statistics, 84 (1): 151–161.

Filippi, V., Ronsmans, C., Campbell, O., Graham, W. J. , Mills, A., Borghi, J., Koblinsky, M., Osrin, D., 2006. Maternal Health in Poor Countries: The Broader Context and a Call for Action. The Lancet, 368: 1535-1541.

Goodburn EA, Chowdhury M, Gazi R, Marshall T, Graham W. 2000. Training Traditional Birth Attendants in Clean Delivery Does Not Prevent Postpartum Infection. Health Policy and Planning, 15(4): 394-399.

Griffiths, P., Stephenson, R., 2001. Understanding Users' Perspectives of Barriers to Maternal Health Care Use in Maharashtra, India. J. biosoc. Sci., 33: 339–359.

Gupta M. D. 1990. Death Clustering, Mothers' Education and the Determinants of Child Mortality in Rural Punjab, India. Population Studies: A Journal of Demography, 44 (3): 489-505.

Heckman, J., Ichimura, H., Todd, P., 1998. Matching as an Econometric Evaluation Estimator. Review of Economic Studies, 65 (2): 261–294.

Heinrich C., A. Maffioli, G. Vazquez. 2010. A Primer for Applying Propensity Score Matching, Impact – Evaluation Guidelines. Inter – American Development Bank.

Hogan, M.C., K. J. Foreman, M. Naghavi, S. Y. Ahn, M. Wang, S.M. Makela., A. D. Lopez, R. Lozano, C. J.L. Murray. 2010. Maternal Mortality for 181 Countries, 1980—2008: A Systematic Analysis of Progress towards Millennium Development Goal 5. The Lancet, 375(9726): 1609 – 1623.

Huber M., M. Lechner, C. Wunsch. 2013. The Performance of Estimators based on the Propensity Scores. Journal of Econometrics, 175(1): 1- 12.

International Institute for Population Sciences (IIPS) and Macro International. 2007. National Family Health Survey, India, 2005 – 06, Key Findings. http://www.measuredhs.com/pubs/pdf/SR128/SR128.pdf, accessed March 2013.

International Institute for Population Sciences (IIPS) and Macro International. 2008a. National Family Health Survey (NFHS-3), India, 2005 – 06: Bihar. http://www.rchiips.org/NFHS/NFHS-3%20Data/Bihar_report.pdf, accessed March 2013.

International Institute for Population Sciences (IIPS) and Macro International. 2008b. National Family Health Survey (NFHS-3), India, 2005 – 06: Gujarat. http://www.rchiips.org/nfhs/NFHS-3%20Data/gujarat_state_report_for_website.pdf, accessed March 2013.

Imbens G., D. Rubin. 2010. Chapter 20 - Causal Inference in Statistics and Social Sciences. http://www.ics.uci.edu/~sternh/courses/265/imbensrubin20.pdf, accessed March 2013.

Jolliffe D. 2002. Whose Education Matters in the Determination of Household Income? Evidence from a Developing Country. Economic Development and Cultural Change, 50(2): 287-312.

Koblinsky M.A. 1995. Beyond Maternal Mortality Magnitude, Interrelationship, and Consequences of Women's Health, Pregnancy-Related Complications And Nutritional Status On Pregnancy Outcomes. International Journal of Gynecology & Obstetrics, 48: 21- 32.

Koblinsky M.A., M. E. Chowdhury, A. Moran, C. Ronsmans. 2012. Maternal Morbidity and Disability and Their Consequences: Neglected Agenda in Maternal Health. J. Health. Population and Nutrition, 30(2):124-130.

Leuven E., B. Sianesi. 2003. PSMATCH2: Stata Module to Perform Full Mahalanobis and Propensity Score Matching, Common Support Graphing, and Covariate Imbalance Testing. http://ideas.repec.org/c/boc/bocode/s432001.html

McLaren L. 2007. Socioeconomic Status and Obesity. Epidemiologic Reviews, 29: 29 – 48.

Mavalankar D. V., C. R. Trivedi, R. H. Gray. 1991. Levels and Risk Factors for Perinatal Mortality in Ahmedabad, India. Bull World Health Organization, 69(4): 435–442.

Navaneetham K., Dharmalingam, A., 2002. Utilization of Maternal Health Care Services in Southern India. Social Science & Medicine, 55: 1849-1869.

National Institute of Health, National Heart, Lung and Blood Institute. 1998. Clinical Guidelines on the Identification. Evaluation and Treatment of Overweight and Obesity in Adults: the Evidence Report. Obesity Research. 6(2): S51 – 210.

Padmadas, S.S., Kumar S., Nair, S.B., Kumari A.K., 2002. Caesarean Section Delivery in Kerala, India: Evidence from a National Family Health Survey. Social Science & Medicine, 51: 511-521.

Rosenbaum, P.R., Rubin, D.B., 1983. The Central Role of the Propensity Score in Observational Studies for Causal Effects. Biometrika, 70 (1): 41-55.

Schaider J., S. Ngonyani, S. Tomlin, R. Rydman, R. Roberts. 1999. International Maternal Mortality Reduction: Outcome of Traditional Birth Attendant Education and Intervention in Angola. Journal of Medical Systems, 23(2): 99 – 105.

Sebire N.J., M. Jolly, J. Harris, L. Regan, S. Robinson. 2001. Is Maternal Underweight Really A Risk Factor for Adverse Pregnancy Outcome? A Population-Based Study in London. British Journal of Obstetrics and Gynaecology, 108(1): 61-66.

Smith, J., P. Todd. 2005. Does Matching Overcome LaLonde's Critique of Nonexperimental Estimators? Journal of Econometrics, 125(1-2): 305 - 353.

Subramanian, S.V., G.V. Smith. 2006. Patterns, Distribution, and Determinants of Under-and - Over Nutrition: A Population-Based Study of Women in India. Am J Clin Nutr, 84 (3): 633-640.

Swait J., J. Louviere. 1993. The Role of the Scale Parameter in the Estimation and Comparison of Multinomial Logit Models. Journal of Marketing Research, 30(3): 305 – 314.

Thind A., A. Mohani, K. Banerjee, F. Hagigi. 2008. Where to Deliver? Analysis of Choice of Delivery Location from a National Survey in India. BMC Public Health, 8: 29.

The World Bank Report. 2005. Towards a Development Strategy – Bihar. http://siteresources.worldbank.org/INTINDIA/Resources/Bihar_report_final_June2005.pdf, accessed March 2013.

Vora K.S., D .V Mavalankar, K.V. Ramani, M. Upadhyay, B. Sharma, S. Iyengar, V. Gupta, K. Iyengar. 2009. Maternal Health Situation in India: A Case Study. Journal of Health Population and Nutrition. 27(2): 184–201.

WHO, UNICEF, UNFPA and The World Bank. 2012. Trends in Maternal Mortality: 1990 to 2010. http://www.unfpa.org/webdav/site/global/shared/documents/publications/2012/Trends_in_maternal_mortality_A4-1.pdf, accessed March 2013.

Table 1: Summary Statistics of Variables Used in Propensity Score Estimation

Categories	**Variables**	**Definitions**	**Means (standard deviations)**			
			Bihar		**Gujarat**	
			1998-99	**2005-06**	**1998- 99**	**2005-06**
Treatment	DELIVERY	1 if last delivery took place at a health facility, 0 if it took place at home	0.14 (0.35)	0.26 (0.44)	0.50 (0.50)	0.57 (0.49)
Outcome	BMISTATUS	1 if Body Mass Index ≥ 18.5, 0 otherwise	0.60 (0.50)	0.55 (0.50)	0.55 (0.49)	0.58 (0.49)
Mother's Characteristics	AGE	Age of mother in years	25.87 (5.85)	27.67 (6.37)	25.11 (5.06)	26.81 (4.96)
	SINGLE	1 if husband staying elsewhere, 0 if not	0.13 (0.34)	0.27 (0.44)	0.02 (0.13)	0.03 (0.16)
	WORKING	1 if mother is working, 0 if not	0.20 (0.40)	0.22 (0.42)	0.36 (0.48)	0.41 (0.49)
Household's Characteristics	SEXHEAD	1 if household head is a male, 0 if female	0.95 (0.21)	0.76 (0.42)	0.94 (0.24)	0.96 (0.20)
	EDUCATION: NONE	1 if minimum education of husband and wife is no education, 0 otherwise	0.76 (0.42)	0.68 (0.47)	0.47 (0.50)	0.39 (0.49)
	EDUCATION: PRIMARY	1 if minimum education of husband and wife is primary education, 0 otherwise	0.07 (0.25)	0.08 (0.27)	0.16 (0.37)	0.16 (0.36)
	EDUCATION: SECONDARY	1 if minimum education of husband and wife is secondary education, 0 otherwise	0.14 (0.35)	0.21 (0.41)	0.26 (0.44)	0.41 (0.49)
	EDUCATION: HIGHER	1 if minimum education of husband and wife is higher education, 0 otherwise	0.03 (0.17)	0.03 (0.16)	0.11 (0.31)	0.04 (0.20)
	WEALTHINDEX: POOREST	1 if the household belongs to the poorest wealth quintile, 0 otherwise	NA	0.29 (0.45)	NA	0.09 (0.29)

	WEALTHINDEX: POORER	1 if the household belongs to the poorer wealth quintile, 0 otherwise	NA	0.28 (0.45)	NA	0.16 (0.36)
	WEALTHINDEX: MIDDLE	1 if the household belongs to the middle wealth quintile, 0 otherwise	NA	0.17 (0.38)	NA	0.19 (0.40)
	WEALTHINDEX: RICHER	1 if the household belongs to the richer wealth quintile, 0 otherwise	NA	0.15 (0.36)	NA	0.25 (0.43)
	WEALTHINDEX: RICHEST	1 if the household belongs to the richest wealth quintile, 0 otherwise	NA	0.11 (0.31)	NA	0.30 (0.46)
	TOILET	1 if household has access to toilet, 0 if not	0.15 (0.35)	0.30 (0.46)	0.39 (0.48)	0.51 (0.50)
	ELECTRICITY	1 if household has electricity connection, 0 if not	0.15 (0.36)	0.33 (0.47)	0.84 (0.36)	0.87 (0.32)
	RESIDENCE	1 if household lives in an urban area, 0 if not	0.08 (0.28)	0.31 (0.46)	0.39 (0.49)	0.40 (0.49)
	RELIGION	1 if religion of household is Hindu, 0 if Muslim	0.82 (0.38)	0.80 (0.40)	0.90 (0.30)	0.88 (0.32)
Child's Birth Order	BORD	Birth order of the last child	3.33 (2.11)	3.79 (2.35)	2.67 (1.69)	2.7 (1.60)
	N	Number of observations	2176	1053	937	868

* The WEALTHINDEX is a consolidate wealth index constructed from 33 household assets. Unfortunately, data on WEALTHINDEX is not available for NFHS 1998 – 99. "NA" stands for not applicable.

Table 2: Logistic Regression Results for Propensity Score Estimation (dependent variable = DELIVERY)

Categories	**Variables**	**Coefficients (standard errors)**			
		Bihar		**Gujarat**	
		1998 – 99	**2005 – 06**	**1998 – 99**	**2005 – 06**
Mother's characteristics	AGE	0.04** (0.02)	0.03 (0.02)	0.06** (0.02)	0.05** (0.02)
	SINGLE	-0.25 (0.21)	-0.21 (0.25)	-0.39 (0.61)	-0.02 (0.50)
	WORKING	-0.68*** (0.24)	-0.14 (0.24)	-0.38** (0.17)	-0.42** (0.17)
Household's characteristics	SEXHEAD	-0.50 (0.31)	-0.23 (0.25)	-0.31 (0.33)	-0.36 (0.44)
	EDUCATION: PRIMARY	0.56** (0.25)	0.17 (0.30)	0.42* (0.21)	-0.18 (0.24)
	EDUCATION: SECONDARY	1.27*** (0.17)	0.60** (0.24)	0.90*** (0.21)	0.28 (0.22)
	EDUCATION: HIGHER	1.84*** (0.33)	1.43** (0.69)	1.82*** (0.39)	1.42* (0.78)
	WEALTHINDEX: POORER	NA	0.24 (0.27)	NA	-0.07 (0.36)
	WEALTHINDEX: MIDDLE	NA	0.40 (0.31)	NA	0.45 (0.37)
	WEALTHINDEX: RICHER	NA	0.57 (0.40)	NA	0.82* (0.43)
	WEALTHINDEX: RICHEST	NA	1.47*** (0.50)	NA	1.35*** (0.49)
	TOILET	0.51*** (0.19)	0.80*** (0.27)	0.46** (0.20)	-0.14 (0.26)
	ELECTRICITY	0.70*** (0.18)	0.19 (0.24)	0.67*** (0.25)	0.34 (0.31)
	RESIDENCE	0.58*** (0.22)	0.68*** (0.21)	0.87*** (0.18)	0.92*** (0.21)
	RELIGION	0.68*** (0.22)	0.77*** (0.26)	-0.35 (0.27)	0.24 (0.25)
Child's Birth Order	BORD	-0.26*** (0.06)	-0.22*** (0.07)	-0.27*** (0.07)	-0.29*** (0.07)
	CONSTANT	-2.84*** (0.54)	-2.73*** (0.61)	-1.51** (0.64)	-1.02 (0.68)

Significance levels: *** 0.01, **0.05, * 0.10. "NA" stands for not applicable.

Table 3: Treatment Effects of Location of Delivery on BMISTATUS of Mothers

Matching Method	**Average Treatment Effect on the Treated (ATT) (standard errors)**			
	Bihar		**Gujarat**	
	1998 – 99	**2005 – 06**	**1998 – 99**	**2005 – 06**
Nearest Neighbor Matching *N = 1*	0.02 (0.05)	0.16*** (0.07)	-0.04 (0.10)	-0.002 (0.06)
Nearest Neighbor Matching *N = 4*	-0.017 (0.04)	0.16*** (0.05)	-0.06 (0.07)	-0.03 (0.05)
Caliper Matching *Tol = 0.01*	0.003 (0.05)	0.16*** (0.06)	0.002 (0.05)	-0.05 (0.05)
Caliper Matching *Tol = 0.05*	0.003 (0.05)	0.16*** (0.06)	0.002 (0.05)	-0.05 (0.06)

Significance levels: *** 0.01, **0.05, * 0.10.

Table 4: Treatment Effects of Location of Delivery on BMISTATUS of Mothers in Bihar (NFHS 2005-06) using Caliper Matching

Radius (*r*)	Percent of on Support Treated Observations	Average Treatment Effect on the Treated (standard error)
0.0001	19%	0.14 (0.10)
0.001	53%	0.10 (0.07)
0.005	82%	0.16*** (0.06)
0.008	88%	0.17*** (0.06)
0.01	91%	0.17*** (0.06)
0.02	97%	0.17*** (0.06)
0.05	99%	0.16*** (0.07)

Significance levels: *** 0.01, **0.05, * 0.10.

Figure 1: Distribution of Body Mass Index of Mothers in Bihar

Figure 1a: 1998-99

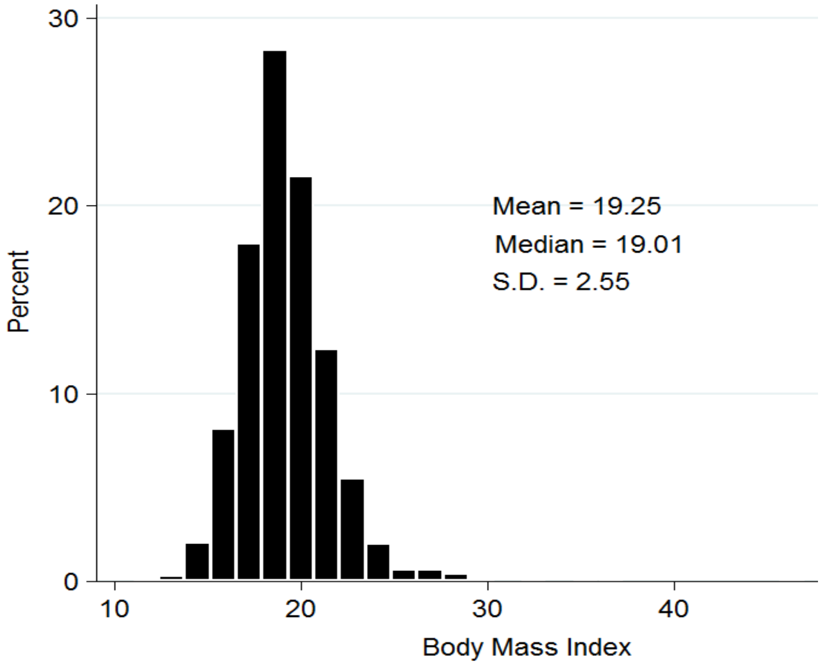


Figure 1b: 2005-06

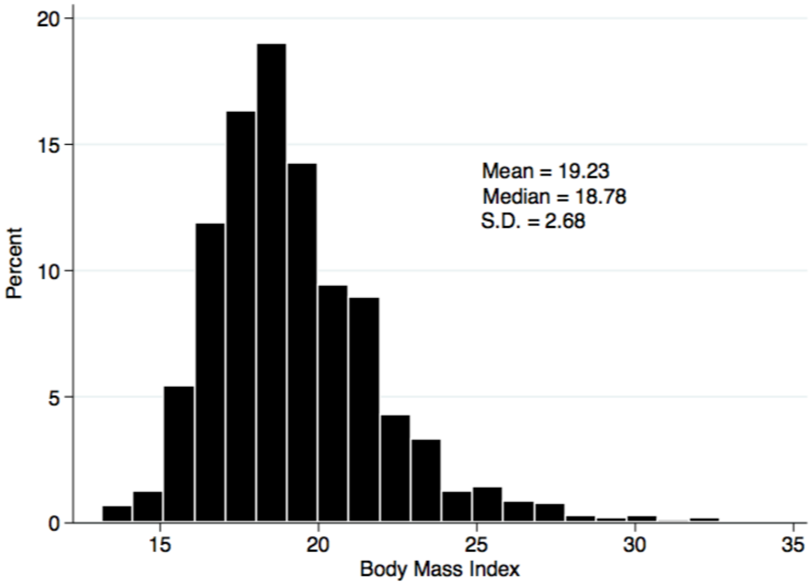


Figure 2: Distribution of Body Mass Index of Mothers in Gujarat

Figure 2a: 1998-98

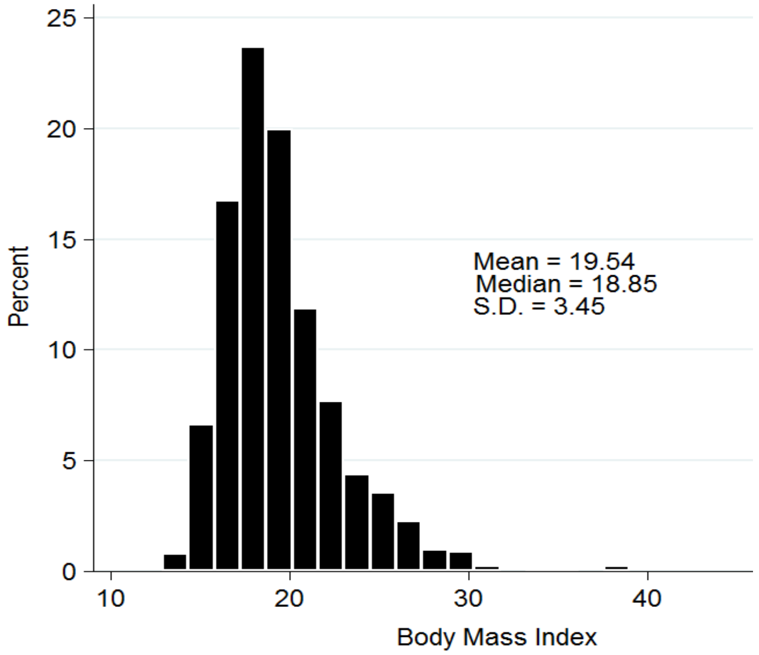


Figure 2b: 2005-06

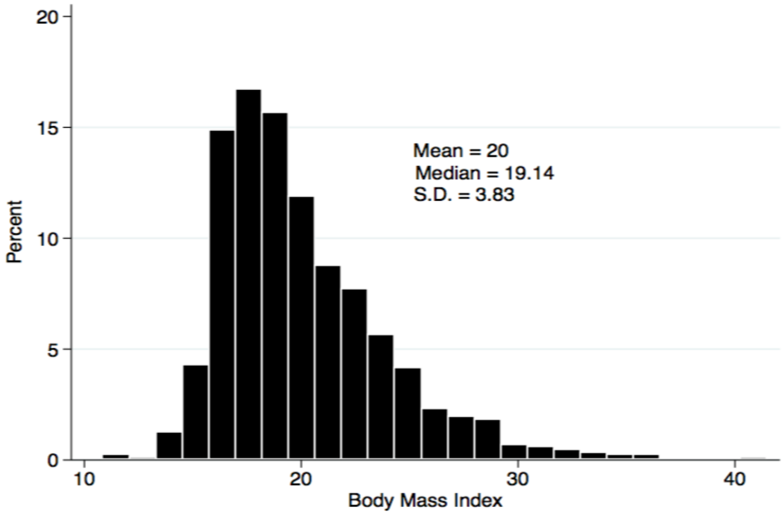


Figure 3: Common Support Conditions for Nearest Neighbor Matching (*N=1*) using NFHS 2005-06 Data

Figure 3a: Bihar

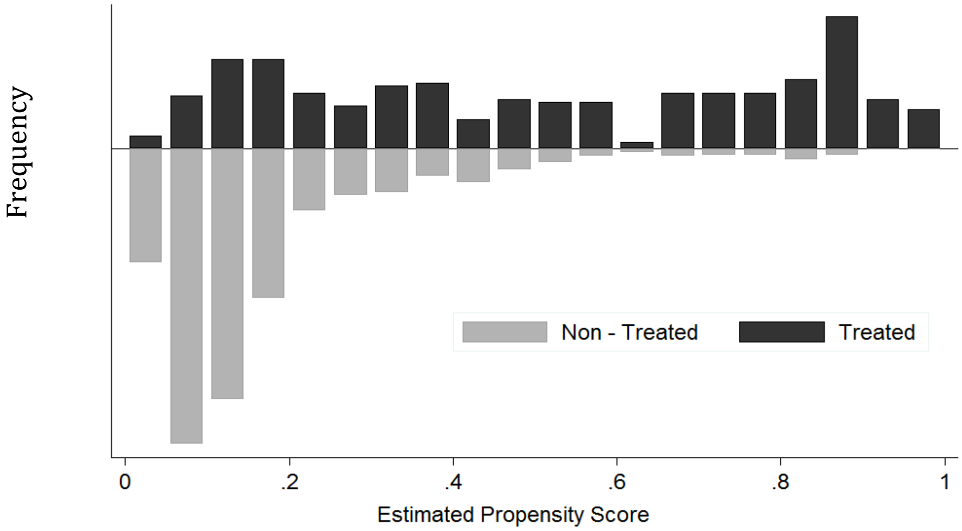


Figure 3b: Gujarat

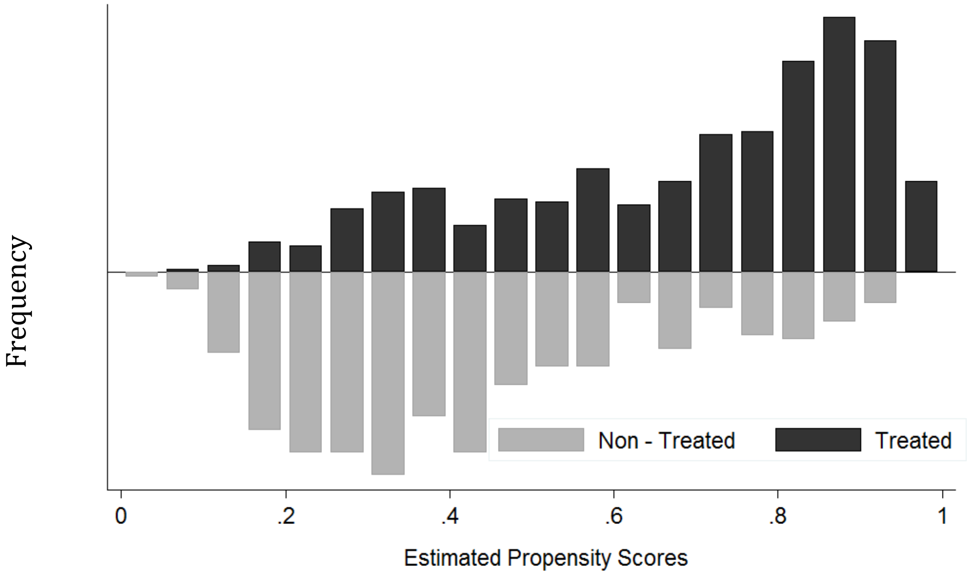


Figure 4: Common Support Conditions for Caliper Matching (*Tol* = *0.01*) using NFHS 2005-06 Data

Figure 4a: Bihar

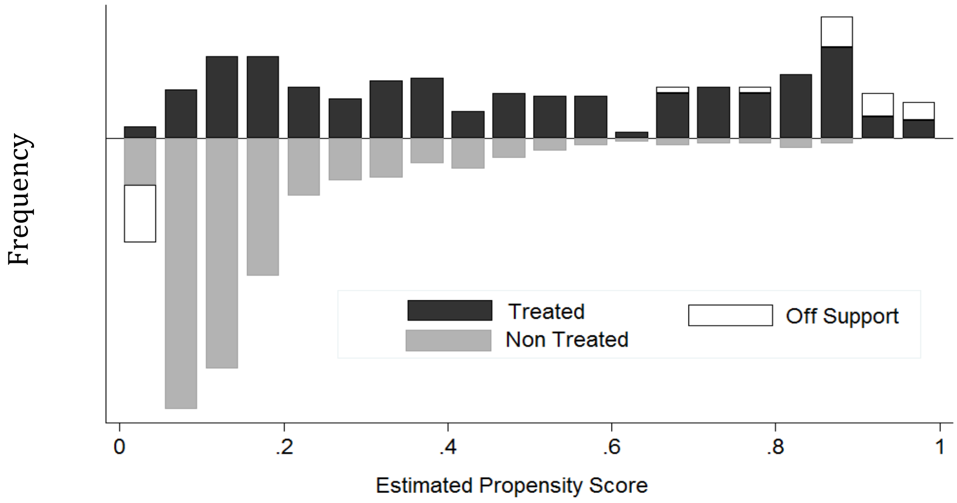


Figure 4b: Gujarat

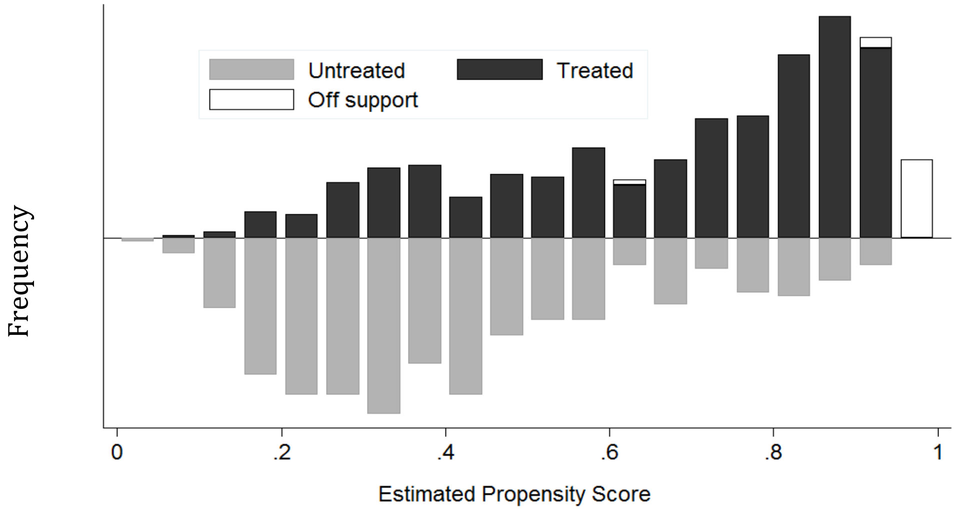


Figure 5: Balancing Condition for Propensity Score Matching via Nearest Neighbor Matching (*N=1*)

Figure 5a: Bihar, NFHS 2005-06

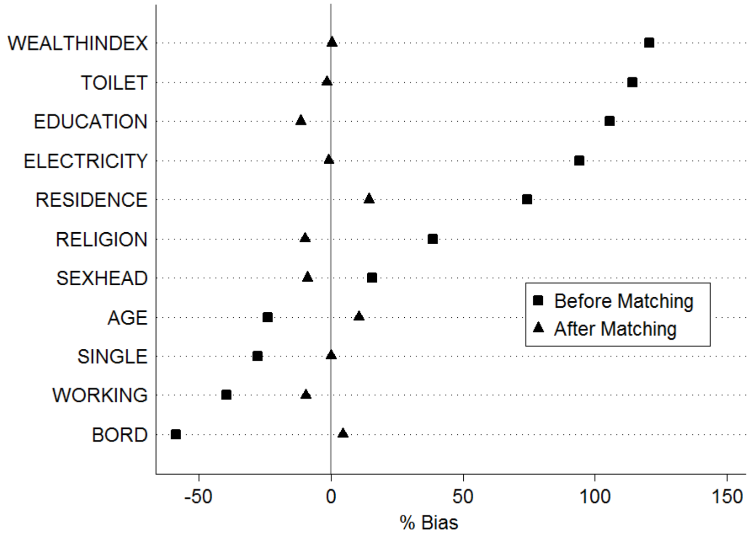
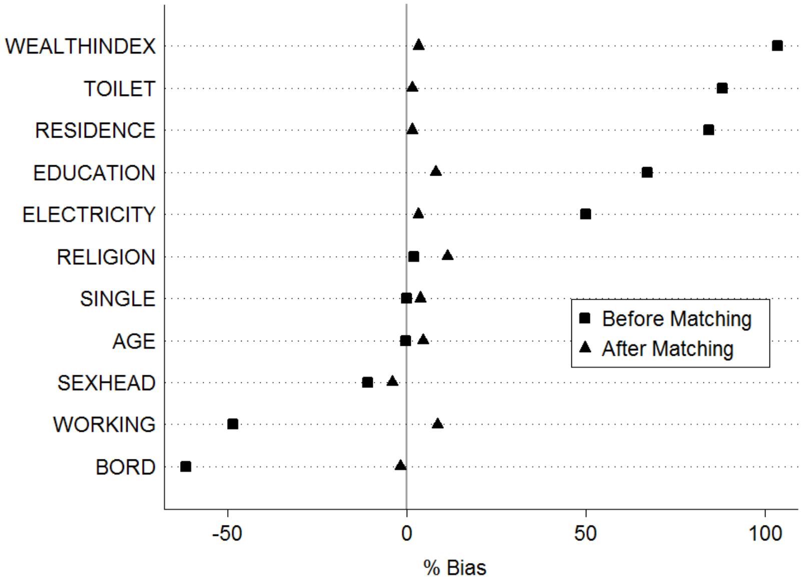


Figure 5b: Gujarat, NFHS 2005-06


WEALTHINDEX
TOILET
RESIDENCE
EDUCATION
ELECTRICITY
RELIGION
SINGLE
AGE
SEXHEAD
WORKING
BORD
■ Before Matching
▲ After Matching
% Bias