

UC Berkeley

UC Berkeley Electronic Theses and Dissertations

Title

Essays in Methodology

Permalink

<https://escholarship.org/uc/item/3dg8v214>

Author

Urbancic, Michael Benjamin

Publication Date

2012

Peer reviewed|Thesis/dissertation

Essays in Methodology

by

Michael Benjamin Urbancic

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Economics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Ulrike Malmendier, Chair

Professor Stefano DellaVigna

Assistant Professor Ming Hsu

Fall 2012

Essays in Methodology

Copyright © 2012

by

Michael Benjamin Urbancic

Abstract

Essays in Methodology

by

Michael Benjamin Urbancic

Doctor of Philosophy in Economics

University of California, Berkeley

Professor Ulrike Malmendier, Chair

Academic contributions in any discipline are only as convincing as the methods used to establish them. This dissertation highlights two methodological issues in economics—one in experimental economics and one in applied econometrics—and argues for increased caution in both the design and the interpretation of empirical studies. Within experimental economics the Becker-DeGroot-Marschak (BDM) mechanism is widely used to elicit the valuations of experimental subjects. Although it is theoretically incentive compatible, empirical evidence suggests that elicitation are affected by the distribution from which the random price is drawn. The second chapter presents a novel, within-subjects data of sequential BDM rounds with varied distributions to directly investigate and characterize distributional dependence. When analyzing data collected outside of the realm of randomized experiments (in the laboratory or otherwise), fixed effects are frequently used to “control for” the potential influence of observed factors on an outcome variable of interest. The third chapter discusses potential pitfalls in the use and interpretation of fixed effects. The goal of each of these chapters is to offer positive suggestions for more careful future research through marginal improvements in empirical design and practice.

Acknowledgements

I am grateful to my advisor, Ulrike Malmendier, as well as to Botond Kőszegi, Shachar Kariv, Stefano DellaVigna, and Matthew Rabin for guidance, encouragement, and support throughout this process. I thank Dan Acland, Michael Anderson, Alan Auerbach, Rodney Andrews, Joshua Angrist, Marianne Bitler, Henning Bohn, Henry Brady, Moshe Buchinsky, Colin Cameron, Carlos Dobkin, Mitch Hoffman, Ed Johnson, Maximilian Kasy, Patrick Kline, Yolanda Kodrzycki, Maciej Kotowski, Trevon Logan, Fernando Lozano, Robert MacCoun, Doug Miller, Juan Carlos Montoy, Enrico Moretti, Ron Oaxaca, Antonio Rangel, Steve Raphael, Jesse Shapiro, Jasjeet Sekhon, Todd Sorensen, Doug Steigerwald, Josh Tasoff, Rocio Titiunik, and Philippe Wingender as well as seminar participants at UC Berkeley, the 2008 AEA Pipeline Conference at UCSB, and the 2009 All UC Labor Conference for valuable comments and suggestions. The BDM experiment would not have been possible without the technical and moral support of Vinci Chow. I thank Rowilma Balza del Castillo and the dedicated staff at the XLab for assistance with the logistics of the experimental sessions. Funding for the project was generously provided by an XLab Pilot Grant. Any and all errors are mine. I am indebted to my coauthors Charlie Gibbons and Juan Carlos Suárez Serrato for their insights, effort, and persistence. Typesetting of this current document was made possible through the help of Nicole Johnson. Finally, I would like to thank my wife and children for their support, patience, and love.

For Maile, Ellie, Benji, and Kaia.

Contents

1	Introduction	1
2	Testing Distributional Dependence in the Becker-DeGroot-Marschak Mechanism	6
2.1	Introduction	6
2.2	Background and Framework	9
2.3	Empirical Study	14
2.4	Descriptive Statistics and Results	19
2.5	Discussion	31
2.6	Conclusion	39
	Appendix 2.A Experimental Instructions and Materials	62
	Appendix 2.B Visualizations of Experimental Data	69
	Appendix 2.C Subjects' <i>ex post</i> Explanations of Their Submissions	81
3	Broken or Fixed Effects?	84
3.1	Introduction	84
3.2	Incorporating heterogeneous treatment effects	86
3.3	Interpreting FE estimates using projection results	87
3.4	A Case Study: Karlan and Zinman (2008)	91
3.5	Fixed Effects Interactions: An <i>AER</i> Investigation	93
3.6	Conclusion	101
	Appendix 3.A Topics in Fixed Effects Theory	103
	Appendix 3.B <code>GSSUtest.ado</code>	108
	Appendix 3.C <i>AER</i> Replications	108
	Bibliography	121

Chapter 1

Introduction

Academic contributions in any discipline are only as convincing as the methods used to establish them. The corollary of this is that scientific progress can only advance as far as its prevailing methods allow.

Consider the role of observation in the realm of celestial mechanics.¹ Astronomers throughout the ancient world managed to assemble an impressive wealth of knowledge about the motions of the Sun, Moon, and planets with respect to the background stars based on systematic naked-eye observations alone. The Egyptians developed a strikingly accurate solar calendar and were able to use their knowledge to successfully predict the annual floods of Nile (which corresponded with rainy seasons far to the south). Aristarchus of Samos estimated the relative distances from the Earth to the Sun and to the Moon by judging the angle between those two bodies at first- and third-quarter lunar phases (although he was off by more than an order of magnitude). Hipparchos used the size of Earth's shadow during lunar eclipses and Eratosthenes' previous estimate of the size of the Earth to calculate the distance to the Moon to within 1% of the true value. More impressively, reviewing records that were ancient even in his day Hipparchos deduced the precession of equinoxes (caused by the Earth wobbling on its axis) and was able to demonstrate the effect over the relatively short time scale of several years.²

¹Examples of the power of methods to enable new lines of empirical investigation and to shape the course of theoretical work abound in any number of disciplines. I choose these examples in celestial mechanics because they are accessible, elegant, and fascinating.

²The cycle of the precession takes approximately 26,000 years to complete, making this early

Nevertheless, the observational methods and tools available to the ancients were too imprecise to distinguish between the two competing models of the Solar System. Aristarchus had argued for a heliocentric arrangement, and he placed the planets in their correct order from the Sun.³ Centuries later, Ptolemy favored an intricate geocentric model in which the planets, Sun, and Moon moved in epicycles centered on perfectly circular orbits around the Earth. Given the data at hand, the two models were observationally equivalent. Actually, the two models differed in one key respect: if the Sun really were at the center and the Earth moved around it, the relative positions of the background stars would be expected to change slightly over the course of the year. This effect—stellar parallax—was eagerly searched for for millenia but never detected. The geocentric model’s prediction of no stellar parallax thus seemed to be a better fit for the empirical facts. Ptolemy’s more developed and sophisticated model was able to make very precise predictions about eclipses and planetary conjunctions, and the geocentric paradigm remained the scientific consensus for 1400 more years.

The painstaking work of Tycho Brahe, whose superior instruments, techniques, dedication, and skill enabled him to plot positions of the planets and stars to an accuracy within 1 or 2 arcminutes (i.e., one-sixtieth or one-thirtieth of a degree, respectively) paved the way for a revolution in celestial mechanics.⁴ Tycho’s meticulous observations led him to rebut Copernicus’ revival of the heliocentric model. Instead, he proposed a hybrid: that the planets in the heavens indeed circle the Sun on circular epicycles centered on circular orbits, but that the Sun (along with its attendant planets) in turn orbits the Earth. In working through the data after Tycho’s death his assistant, Johannes Kepler, was troubled that neither Tycho’s model nor that of Copernicus (which also used circular orbits and epicycles) could be completely reconciled to the observed positions of Mars. With discrepancies of up to 8 arcminutes—well outside of Tycho’s presumed measurement errors—Kepler sought

achievement truly remarkable.

³See Heath (1913) for a thorough account of this forgotten forerunner of the celebrated Copernicus, who lived in the third century BC.

⁴Tycho had set for himself the goal of being within 1 arcminute of accuracy in his measurements, but he often fell short of this mark (especially with dimmer stars). His median accuracy has later been determined to have been 1.5 arcminutes, with a mean accuracy of 2 arcminutes. Rawlins (1993) provides a detailed discussion on Tycho’s 1004-star catalog.

alternative models to fit the data. After trying dozens of ovoid-like orbital shapes in vain, Kepler found that the data fit an elliptical orbit perfectly.⁵ This revolutionary insight formed the first of Kepler’s three laws of planetary motion, which would not have been possible were it not for Tycho’s unprecedentedly accurate observational data.

Although Kepler’s laws of planetary motion—later explained by Newton’s laws and the attractive force of gravity—firmly put the Sun at the center of the Solar System, it wouldn’t be until the nineteenth century for observations to be sufficiently accurate to measure stellar parallax. This feat required measurements that were precise to small fractions of an arcsecond: two orders of magnitude more precision than Tycho’s visual observations. To be able to do this, the astronomers of the late 1830s required technical advances in optics and refractive telescopes, and they needed to account not only for the mechanical idiosyncracies of their instruments over the course of several months (or even years) but also myriad effects stemming from the Earth’s atmosphere and its motions.⁶ Aided by his own considerable contributions to the methods of “reduction” (i.e., correcting for the various effects of the Earth’s atmosphere and motion so that observations across instruments, places, and times could be legitimately compared), Friedrich Bessel was able to publish a successful parallax measurement for 61 Cygni (Piazzi’s “Flying Star”) in 1838.⁷ Again, slow and painstaking advances in measurement methods were fundamental to this monumental achievement.

⁵Ironically, he had initially skipped an elliptical model in the belief that it was so simple that someone must have previously tried it. See Caspar (1959).

⁶This exacting degree of accuracy required correction for *atmospheric refraction* (in which light passing through the atmosphere is bent differentially depending on the angle at which it enters, causing star’s apparent positions to depend on their elevation in the sky), the *precession* of the Earth’s axis, the *nutation* of the Earth’s axis (a swaying to and fro perpendicular to the precession, with a first-order cycle of 18 years), and *aberration* (a consequence of the finite speed of light and the nontrivial movement of the Earth along its orbit, which causes it to move either slightly toward or away from a background star at six-month intervals). See Clerke (1885).

⁷The culmination of this millenia-long quest is put most eloquently by Agnes Mary Clerke in her treatise on nineteenth-century astronomy (Clerke (1885)): “The resulting parallax of 0.3483” (corresponding to a distance about 600,000 times that of the earth from the sun), seemed to be ascertained beyond the possibility of cavil, and is memorable as the first *published* instance of the fathom-line, so industriously thrown into celestial space, having really and indubitably *touched bottom.*” (Emphasis hers.)

This dissertation consists of two methodological papers that focus on issues of measurement within the discipline of economics. The first investigates the possibility that the Becker-DeGroot-Marschak mechanism (widely used in experimental economics for measuring a subject’s valuation for a good) might in practice exhibit sensitivities not predicted by theory. The second looks at potential pitfalls in the application and interpretation of fixed effects, which are often used in observational research to “control for” the effects of observed variables that might also influence an outcome variable (in addition to a right-hand-side variable of interest). Each of these chapters is infinitely more humble in scope than the examples of Tycho, Bessel, and the work of innumerable other scientists and academics. I cite these giants only to underscore the importance of the task of improving a discipline’s methods—certainly not for the briefest of moments to compare this work to theirs (which comparison would be catastrophically unflattering for this body of work).

Despite its potential importance for empirical and theoretical researchers alike, methodological studies are not carried out often enough in the discipline of economics. One explanation for this underprovision is readily understood by economists: methodological research is likely to exhibit positive externalities, since it offers the promise of benefitting other researchers through improving the quality of their work. As only a small fraction of these benefits will accrue to methodological researchers themselves (through the channel of citations, for instance), the incentives for providing this research are not as attractive as might be collectively optimal.

Unfortunately, rather than following its own policy prescriptions and devising institutions or incentives to overcome the underprovision of a service yielding positive externalities, the discipline all too frequently actively discourages methodological inquiry. Multiple researchers either explicitly discouraged or cautioned against carrying out the BDM project discussed in Chapter 2, citing a lack of career rewards—especially for someone still in graduate school.⁸ As for the fixed-effects study found in Chapter 3, my coauthors and I were advised by another researcher to abandon the project altogether, not because of any failing in our approach or contribution, but

⁸One of these researchers genuinely believes that methodological studies are underprovided, and when I reaffirmed that I would be moving forward with the study regardless of the consequences he replied: “I’m glad you’re jumping on that grenade.” In other words, it was best for everyone that *someone* did the work, even though things might not go well for that particular someone.

rather because he perceived that it might cause offense to some of the researchers whose work we had replicated and discussed (or to any number of others in a similar position). He was concerned about the possibility of future career repercussions for us and also the reputation of our graduate program.

Would that this state of affairs were reversed! In addition to the marginal advances that may come from the content of these methodological papers, my hope is that the very existence of this dissertation might in some way encourage more talented researchers than I to pursue methodological studies in the future. For as with any other discipline, economics is only as good as the quality and accuracy of its methods. *Per aspera ad astra.*

Chapter 2

Testing Distributional Dependence in the Becker-DeGroot-Marschak Mechanism

2.1 Introduction

In laboratory and field experiments in economics and marketing it is often necessary to determine how much subjects (consumers) value a given good. Asking subjects hypothetically how much an item might be worth to them is cheap and easy, but research suggests that respondents tend to overstate their valuations if there is no cost in doing so.¹ Fundamentally, elicitation mechanisms should be incentive compatible: subjects must not stand to benefit from reporting their valuation insincerely.

An ingenious solution to this problem is to decouple the determination of whether or not a transaction takes place from the determination of the transaction price. In the method presented by Becker, DeGroot and Marschak (1964), a subject first states her willingness to pay for an item. Afterward, a number is drawn at random from some distribution, which is usually known to the subject when reporting her valuation. If the stated willingness to pay is less than the randomly drawn number, the subject does not receive the item, and she pays nothing. If the stated willingness

¹For more on the existence of “hypothetical bias” see List and Gallet (2001).

to pay is greater than or equal to the randomly drawn number, the subject receives the item and pays a price equal to the drawn number—not the stated willingness to pay.

Faced with the Becker-DeGroot-Marschak (BDM) mechanism, an expected-utility-maximizing subject has a weakly dominant strategy to state her valuation truthfully.² If she *understates* her valuation and the randomly drawn number falls between her stated and sincere valuations, she will miss the opportunity to obtain the good for a price she would have been willing to pay, thus forfeiting surplus. If she *overstates* her valuation and the randomly drawn number falls between her stated and sincere valuations, she will obtain the good but for a price higher than she would like to pay, thus incurring negative surplus. Moreover, reporting her valuation sincerely remains a weakly dominant strategy regardless of her attitude toward risk. Due to its incentive compatibility and relative ease of implementation, the BDM mechanism has been used in hundreds of empirical studies to measure subject valuations.³ The BDM mechanism is chiefly associated with experimental economics, but it has also been used in psychology and marketing experiments.⁴

In theory, the BDM mechanism is incentive compatible regardless of the distribution from which the random price is drawn.⁵ Whether the distribution is uniform, normal, triangular, *et cetera*, and no matter where the subject's valuation is thought

²If the valuation falls outside of the support of the distribution (either outside its extremes or between two discrete mass points), local deviations may have no effect on the outcome, which is why the strategy of reporting one's valuation sincerely is weakly rather than strictly dominant.

³As of November 2011, Google Scholar lists 869 publications and papers that cite Becker, DeGroot and Marschak (1964). The five most widely cited of these are Kahneman, Knetsch and Thaler (1990) with 1962 citations, Glaeser et al. (2000) with 1177, Machina (1987) with 960, Grether and Plott (1979) with 906, and Bateman et al. (2002) with 746.

⁴The BDM mechanism was a key element in the studies on the preference reversal phenomenon, which was first reported by experimental psychologists Lichtenstein and Slovic (1971). For examples of the uses of the BDM mechanism in marketing, see Hoffman et al. (1993) and Wertenbroch and Skiera (2002).

⁵A number of researchers have raised concerns about the incentive compatibility of the BDM mechanism for goods that aren't concrete. For instance, Karni and Safra (1987); Safra, Segal and Spivak (1990); and Keller, Segal and Wang (1993) demonstrate and explore why the BDM mechanism is not necessarily incentive compatible if the item on offer is a lottery. The present study, however, focuses on privately valued concrete goods. Harrison (1992) casts doubt on the usefulness of the BDM mechanism in lottery settings for payoff dominance reasons: the expected costs from deviating from reporting one's sincere valuation may be trivially small.

to lie with respect to the support, the reasoning above holds: any deviation from stating one's sincere valuation may lead to ex post regret either from failing to obtain an item for an attractive price or from obtaining an item for an excessive price. Despite this strong theoretical result, there is suggestive evidence that even for concrete goods responses to the BDM mechanism are not in practice independent of distributions of random prices used.

This chapter reviews the existing suggestive evidence for distributional dependence and posit a framework for understanding the observed patterns in those data. I discuss existing theories for distributional dependence as well as their testable implications. I design a novel, within-subjects laboratory experiment with varied distributions across sequential BDM elicitation rounds to test the relevant predictions of these models. This design allows me to investigate the role of learning by the subjects over the course of the experiment. By giving half of the subjects more detailed instructions, I can also look at how responses differ by their level of comprehension of the institution.

I find that the responses of nearly all subjects are sensitive to the distributions to at least some degree and that this sensitivity is significant for roughly half of the subjects. Nearly all of the subjects who are sensitive to distributions manifest mass-seeking bias: the reports of their valuations are systematically moved from their average elicitation in the direction of the mean of the distribution for a given BDM elicitation round. Moreover, for the subjects who are responsive to the distribution of random prices used, the data suggest that this mass-seeking bias is diminishing in the distance of the mass from the subjects' average elicitations.

While the magnitude of this dependence decreases over the course of the sequential rounds—interpreted as the effect of learning from the information in the set of distributions—the sensitivity persists for a majority of the subjects throughout the length of the experiment. Also, subjects with more detailed instructions regarding the BDM institution are markedly less sensitive to these distributional effects, but survey responses suggest that much of this increased stability may be a result of systematic underreports relative to sincere valuations.

The remainder of this chapter is organized as follows. Section 2.2 offers a framework to account for the observed patterns, reviews existing theoretical explanations for distributional dependence in the BDM mechanism, and presents testable hypothe-

ses to distinguish their relative merits. Section 2.3 presents the experimental design. Section 2.4 presents summarizes the data and presents the results. Section 2.5 discusses possible mitigations of distributional dependence and explores implications for future research. Section 2.6 concludes.

2.2 Background and Framework

Two studies serve to illustrate the phenomenon of distributional dependence in the BDM mechanism. First, Bohm, Lindén and Sonnegård (1997) endow subjects with a card redeemable for 30L of gasoline and use the BDM mechanism to ascertain the minimum price they would be willing to accept (WTA) to sell the card back to the experimenters. Using uniform distributions with differing supports for the randomly drawn numbers (each distribution representing a distinct treatment), they find that a higher explicit upper bound on the distribution of random prices leads to a significantly higher average reported WTA.

Second, Mazar, Kőszegi and Ariely (2009) use the BDM mechanism to elicit subjects' maximum willingness to pay for a concrete good (a mug or box of chocolates, for examples). In this study they vary the shapes—though not the support—of the distributions of the randomly drawn numbers across treatment group and find in some of their experiments that treatment groups that faced distributions with more mass to the right have significantly higher average elicitation.

At this time there is no widely accepted theory of why BDM elicitation might be sensitive to the distribution of random draws used. As seen in Section 2.2.2 below, some of the existing models that attempt to explain the phenomenon make predictions that do not seem to match the patterns observed in the studies of Bohm, *et al.* and Mazar, *et al.*, which both suggest mass-seeking bias in the responses of subjects.

2.2.1 Proposed Framework and Testable Hypotheses

Possible explanations for mass-seeking bias in BDM elicitation include rationally inferring of information, misunderstanding of the implications of the institution, gaining utility from making the outcome of the random draw more uncertain and thus more

exciting, finding reassurance or satisfaction in avoiding extremes, and (in some cases) mistakenly approaching the BDM mechanism as if it were analogous to a first-price auction. A subject's perceptions about her valuation—what it is and how that compares to what it should be—is a key piece in each of these explanations. These perceptions may be profoundly affected by the position of the subject's valuation relative to the range of the support of the possible prices and the distribution of mass over that support.

Regardless of the ultimate causes of the mass-seeking bias observed by Bohm, *et al.* and Mazar, *et al.*, the observed patterns suggest the following framework. Let $f(x)$ be the probability distribution [or density] function for distribution of random prices for the BDM mechanism in question, and assume that $f(x)$ has a finite support from a to b . Let v represent the subject's true valuation and \hat{v} be the subject's stated valuation.

First, if $f(\cdot)$ is uniform \hat{v} is expected to be greater than v when more of the support of $f(\cdot)$ is above v than below v . Second, regardless of the shape of $f(\cdot)$, \hat{v} is expected to be greater than v when more of the mass of $f(\cdot)$ is above v than below v . Third, it is expected that there is diminishing sensitivity: the farther a given support or amount of mass is from v , the less pronounced its attractive effect will be on \hat{v} .

Many functional forms may of course be contemplated to implement the broad principles outlined in the framework. The range of the support and the relative positions of the mass may interact in a variety of ways. Note, in particular, that there may be places on some distributions where the first and second points of the framework would work in opposing directions. *A priori* there is no theoretical basis to support any specific parametric formulation. Indeed, there may be multiple correct functional forms given the possibility of heterogeneity in how subjects approach the institution. An agnostic approach is warranted at this stage.

Motivated by the framework discussed above and previous work (more on this in Section 2.2.2), the following specific null hypotheses may be tested:

Proposition 1. (*Dependence*): *Elicited valuations for an identical item are the same—within subject—regardless of the distribution of the randomly drawn prices faced by the subject.*

Proposition 2. (*Mass*): *Relative to a given distribution, subjects do not report higher valuations when faced with a distribution where the mass is located farther to the right (in the sense that the second distribution first-order stochastically dominates the given distribution).*

Proposition 3. (*Comprehension*): *More detailed instructions (H3a) do not affect the magnitude of subjects' elicitation, and they (H3b) do not decrease the variability of subjects' elicitation.*

Propositions 1 and 3b—as well as any effects from learning—cannot be tested with across-subjects data from one-shot BDM elicitation for a given item, which has been the predominant design employed in studies that implement BDM elicitation. Section 2.3 presents the experimental design for obtaining a novel, within-subject data set with varied distributions across rounds to test these hypotheses.

2.2.2 Predictions of Existing Models

A few papers have explored models that imply distributional dependence in the BDM mechanism, but they have the drawback that they often predict that agents will have mass-fleeing bias, which is contrary to the evidence presented in Bohm, et al. and Mazar, et al. Following is a brief summary of three models, together with the predictions that these models would make with respect to the hypotheses presented above given an experimental setting with multiple BDM rounds with varied distributions.

Horowitz (2006) argues that distributional dependence could be the result of asymmetric regret. If different weights are placed on the negative outcomes of not obtaining the item for a price less than one's true valuation and obtaining the item for a price higher than one's true valuation, an agent minimizing ex post regret might exhibit sensitivity to the distribution of random prices.

The implications of the model of Horowitz are broadly consistent with an agent facing probabilistic outcomes who has expectations-based reference-dependent utility, as in Köszegi and Rabin (2006) and Köszegi and Rabin (2007). Having a higher valuation relative to a given distribution of random prices would raise the reference point for obtaining the item, in turn increasing the agent's stated willingness to pay

to avoid incurring a more painful loss. This phenomenon is called the attachment effect by Kőszegi and Rabin.

If a subject with a moderate and certain valuation experienced asymmetric regret according to the model of Horowitz, her bids would be biased upward (downward) when facing distributions with the mass primarily below (above) her valuation.⁶ That is, her bids would be expected to be higher when the mass of the distribution was lower and *vice versa*.⁷ Note that the random selection of one round for resolution should theoretically remove any incentive for hedging across rounds; as long as the resolution of the chosen round were transparent and salient, there would still be ample scope for ex post regret of either type.

Lusk, Alexander and Rousu (2007) argue that when agents faced with the BDM mechanism experience some uncertainty about their true valuation they will have an incentive to deviate from the point estimate of their valuation in the direction of lower expected cost. If the distribution of random prices is locally asymmetric, an agent unsure of her exact valuation will tend to overstate it if there is less mass above her estimate of her valuation and understate it if there is less mass below this estimate. This explanation is similar to that of Horowitz: subjects minimize regret by choosing to err in the direction of lower expected cost.

Both within and across subjects, the Lusk, *et al.* model implies that the triangular distributions should see more dramatic biasing effects than the other distributions, since they present subjects with far more possibilities for local asymmetry of mass around their point estimates of their valuations.⁸ Like that of Horowitz, this model

⁶Insofar as these effects might intensify with respect to the extremity of a given valuation, slight adjustments still might be discernable even when the valuation is always on one side of the median of the various distributions. However, these would be second-order adjustments. Subjects with very low or very high valuations would not vary their bids dramatically across distributions in the model of Horowitz.

⁷One important caveat: if the magnitude of any bias from asymmetric regret were too small relative to the level of discreteness of the distributions used in this experiment the last bids (at least) should be identical across all distributions (within subject).

⁸For example, if a subject's estimate of her valuation for the item in this experiment were \$6 plus or minus \$2, this model would predict that her last bids (at least) would be the same for two distributions that were flat from the entire support from \$0 to \$9.50, save for a spike at either \$0 or \$9.50 (i.e., distributions #3 and #4—see Table 2.2.1). However, triangular distributions (such as distributions #11, #12, #13, and #14—see Table 2.2.1) would yield differences even in the last bids. (That is, unless the subject's estimates of her valuation were at the extremes or beyond the

predicts that across subjects the effect of varying distributions (apart from any effects from learning) should be to push elicitation *away* from mass.

Kaas and Ruprecht (2006) also explain that distribution dependence may occur when agents experience uncertainty about their true valuations, which they conceptualize as “discriminal dispersion”: agents perceive value with a normal distribution of errors centered on their true valuation.⁹ The implications of the Kaas and Ruprecht model are analogous to a specific case of the model of Horowitz in which agents nearly always place more weight on the potential regret of paying too much for an item. However, unlike Horowitz or Lusk et al., Kaas and Ruprecht come to the conclusion that due to the asymmetry of the potential for loss from overbidding, risk-averse agents will nearly always underbid when faced with a BDM elicitation task.

The model of Kaas and Ruprecht, expects most subjects to underbid regardless of their estimate of their valuation or the distribution of random prices. The degree of underbidding is expected to increase in the amount of mass of the distribution to the right of the estimate of the valuation.¹⁰ While direct measures of whether or not underbidding has occurred cannot be obtained from the BDM phase data alone, later survey questions asking about the subjects’ valuations for the item may provide some guidance on the matter (albeit admittedly imprecise). The model of Kaas and Ruprecht also predicts that across subjects underbidding will be reduced (or even reversed for high valuations) for those subjects who are less risk averse.

In contrast, the standard model would predict that a rational expected-utility maximizer with a fully formed and certain personal valuation of the item from the beginning of the experiment would simply bid precisely that valuation in every round of the BDM phase, regardless of the level of detail of instructions or the distribution of random prices in any given round.¹¹

support of the distribution, or if her considered range of adjustment were smaller than the intervals between the mass points.)

⁹A key limitation of Kaas and Ruprecht in the context of this study is that their analysis assumes a uniform distribution of random prices throughout.

¹⁰Even if there were generally underbidding for most or all distributions the magnitude of the differential influences toward underbidding *across distributions* would be relatively small (similar in principle to the cases of extreme values in the Horowitz model).

¹¹This agent wouldn’t necessarily bid her certain valuation if it happened to lie above (below) the support of the distribution of random prices for a particular round. In these instances, she would be indifferent over all the prices above (below) the maximum (minimum) of the support.

If subjects are uncertain about their valuations they may learn from the varied distributional information presented in successive rounds. Additional insight about the role of learning can thus be gained if subjects are allowed to revisit and revise bids made in earlier rounds. In this setup, all of these models would predict that if the subject were uncertain about her valuation and learned about the item's value from the distributional information presented, she would adjust her valuation accordingly, and her *last* elicitation for the various rounds would be (weakly) more closely clustered than her *first* elicitation. Note that the last elicitation of a rational expected-utility maximizer from the standard model should be exactly the same across all 20 rounds.

The predictions of the framework presented in section 2.2.1 differ appreciably from those made by the models above. Unlike the standard model (in particular) it predicts that even subjects' last bids may well differ across rounds with different distributions of random prices. Contrary to the models of Horowitz and Lusk, *et al.* it predicts that subjects' bids will be biased *toward* the direction of greater mass or longer range of the remaining support—not just for first bids, which may be in the process of converging by reason of learning, but also for last bids.

The first and second principles of the framework predict that subjects given the same remaining range of the support and the same mass in either direction from the subject's valuation should experience identical bias. For example, if the subject's valuation were \$4, the first and second principles would predict that the seven distributions with high support (from \$5 to \$14.50) would all lead to the same bias and thus the same bid (at least for the last bids).

In contrast, the third principle predicts that the more distant mass is from a subject's valuation, the less bias it promotes in its direction. In the example from the preceding paragraph, a subject with a valuation of \$4 would be expected to bias his response upward more for the high-support distributions with mass closer to \$5, the lower bound of the support. For example, a subject with a valuation of \$4 would be expected to bid weakly more when faced with distribution #5 (with a spike at \$5) than when faced with #6 (with a spike at \$14.50).

2.3 Empirical Study

The ideal data to test the question of distributional dependence of the BDM mechanism would consist of multiple elicitations of a subject's valuation of a fixed good when the subject is faced with a variety of distributions simultaneously but in isolated parallel. That is, facing each elicitation task at the same moment in time while having no memory of ever seeing any of the other tasks. Alas, this experimental design is—of course—not possible.

The vast majority of studies that rely on the BDM mechanism to measure subject valuations use the mechanism once per item per subject. Moreover, the same distribution—usually uniform over a specified support—is typically used for all subjects for any given item. To my knowledge no empirical study has collected within-subject data for multiple sequential BDM elicitations where subjects face both varying and known distributions.¹²

In February of 2010 I conducted a pilot study using the BDM mechanism in which subjects' valuations were elicited hypothetically for six separate concrete items.¹³ Two different distributions were used for each item. For each item half of the subjects viewed a distribution with more mass to the left first followed by one with more mass to the right (separated by elicitations for other items). The other half encountered the distributions in the opposite order.

Subjects reported significantly higher valuations for every item when faced with more right-massed distributions relative to left-massed distributions. Looking at the second elicitation only—by which time all subjects had seen the same aggregate distributional information—the variance of elicitation across subjects decreased, suggesting the presence of learning. However, for each of the six items the group of

¹²Investigating why the BDM mechanism might not be incentive compatible with lotteries, Keller, Segal and Wang (1993) present subjects with multiple BDM elicitation tasks with uniform distributions with different ranges. In the context of comparing the BDM mechanism with the Vickrey auction, Noussair, Robin and Ruffieux (2004) do ask individual subjects to respond to multiple BDM elicitation with varying distributions, but these distributions are unknown to the subjects and are determined by the responses of another set of subjects rather than by the investigators.

¹³51 subjects participated in the pilot. The six concrete items were a one-quart jar filled with pennies, an unopened deck of black-backed Bicycle playing cards, a pack of Orbit spearmint gum, a 4-GB SD card with built-in USB connectivity, a blue coffee mug with a yellow Cal logo, and an unopened DVD of the movie *WALL-E*.

subjects who faced right-massed distributions second still reported higher valuations than the group of subjects who faced left-massed distributions second, and these across-subjects differences were significant for five of the six goods. The results of the pilot study suggest that Propositions 1 and 2 would be rejected in a larger experiment.¹⁴

2.3.1 Experimental Design

The present, larger experiment took place in late February and early March 2011. It was conducted in a lab so that subjects could physically handle the item on offer, to aid in the credibility of the randomizations, and to facilitate subject payments. On entering the lab each of the subjects sat at an individual computer terminal, the screen of which could not be viewed by any other subject. After brief verbal instructions (which also appeared on the login screen in front of each of them as they waited for the experiment to begin) the subjects were handed a login code and were then able to proceed with the entirety of the remaining experiment individually and at their own pace. Before logging in, subjects were informed that they had been given an endowment of \$15, the amount of which could change during the course of the experiment based on their decisions and on chance.

Nearly all of the experiment was mediated via an internet browser. This allowed for the responses to be prescreened and corrected. For instance, subjects were reminded to submit numeric values when necessary. While subjects encountered the rounds of the BDM phase sequentially, they were allowed to revisit and revise past submissions. The timing of all subject responses was recorded, so the informational history of every subject at each time they submitted a reported valuation (either initially or as a revision) is known.

The item used in this experiment was a gift certificate for a dozen cookies from a new and popular ice cream shop near campus. This item had a relatively modest price, so it could credibly be obtained with the amount of money given to subjects as an endowment at the beginning of the experiment. Since the item was perishable and

¹⁴The same instructions were used for all subjects in the pilot study; its data make no predictions on Proposition 3.

had nearly universal appeal, it was more likely to maximize the number of subjects reporting a positive valuation, thus yielding a richer data set.¹⁵

To reduce the possibility that the subjects might anchor their responses on a numerical cue from a denominational value on the gift certificates, they were simply labeled “DOZEN COOKIES ONLY” on the back. By arrangement with the proprietor of the business, this specially labeled gift certificate was only redeemable for one dozen cookies and could not be used toward any other purchase. The signature item of the shop is a scoop of ice cream served between two freshly baked cookies. A dozen cookies may be bought at once and a price appears for this purchase on the menu, but in practice patrons very rarely make this purchase. Thus while the establishment and its wares would be known to many of the subjects, it was very unlikely that any of them would remember the market value for a gift certificate for a dozen cookies. No subject in any of the sessions inquired about either the value on the gift card or the price of the good on offer.¹⁶ The gift card was handed directly to the subjects individually at the same time as the login code and hard copy of the instructions.¹⁷

Upon logging in subjects privately read one of two possible sets of instructions that explained how the BDM mechanism works. Half of the subjects were given a simple set of instructions for the BDM mechanism, while the other half were given a more detailed set of instructions that explicitly worked through the logic of why submitting one’s true valuation is the optimal strategy. The more detailed instructions were modeled after those used in Plott and Zeiler (2005): subjects in this treatment were explicitly told (in part) that: “Given this set of rules, there is every incentive for you to report your willingness to pay truthfully.” Also, the detailed instructions

¹⁵One significant concern in choosing the item was that a subject who had a high value for a low-cost durable item (say, the traditional school-themed mug or water bottle) would already own one (or more) and thus would not place very much value on obtaining another, while those who did not already own one lacked one precisely because they placed very little—if any—value on it.

¹⁶Several weeks after the experimental sessions were completed, the business in question raised the prices of all of its products, but these gift certificates were still honored and redeemed for one dozen cookies. The author—who retains many of the certificates not obtained by subjects—has made no attempt to exploit the minor arbitrage opportunity thereby presented.

¹⁷Unlike the first and second experimental sessions, the set of seats in the Xlab to be used by the subjects during the third session was predetermined, and one of the gift cards was placed at each of the terminals before the subjects entered the lab. Immediately on seeing the gift card, a few of the first subjects to enter the lab verbally expressed excitement and pleasure.

gave numerical examples of why deviating from the strategy of reporting one’s true valuation could potentially result in *ex post* regret.¹⁸ To prevent subjects from quickly clicking through the instructions, the button to proceed to the next stage of the experiment was locked for three minutes. The subjects were also provided with a hard copy of the instructions that they could reference throughout the remainder of the experiment.¹⁹

The subjects faced two phases of experimental tasks, after which followed the resolution of their payoffs and a brief survey. During the first phase they encountered 20 stages of BDM elicitations with varying distributions, and they were asked to submit their willingness to pay for the item. It was explained to subjects beforehand that exactly one elicitation task was to be selected at random for actual resolution through the BDM mechanism, and the point was explicitly made that they would therefore have an incentive to treat each round separately.

To more cleanly test (across subjects) whether and how responses might change simply when they are elicited more than once, half of the subjects viewed the same distribution in the first two BDM rounds, while the other half viewed that distribution in only the first round and a different distribution—identical across subjects in this treatment—in the second round. The BDM phase thus implemented a 2×2 design over the first two rounds:

		First two rounds	
		AA	AB
Instructions	Simple	SAA	SAB
	Detailed	DAA	DAB

The remaining 18 rounds were randomized for each subject individually. Overall, the same 14 distinct distributions appeared for all subjects over the course of the 20 rounds. Table 2.2.1 lists the full probability density function as well as the mean and

¹⁸The specific additions made to the detailed instructions are shown in the boxed areas on the second page of subject instructions found in Appendix 2.A.

¹⁹The text and images in the hard copies matched exactly the instructions viewed on screen for each subject. The two different versions of the instructions—basic and detailed—had previously been matched and physically attached to the appropriate login codes. The first sheet of the hard copies was indistinguishable, so subjects had no way to know that there were multiple versions of the instructions.

median for each of the 14 distributions used. The subjects did not observe any of the distributions until they encountered them during the course of the BDM phase, thus some earlier than others. A predetermined set of six of the distributions were repeated exactly once each at some point for each subject (including the distribution repeated in the first two rounds for the subjects in treatments SAA and DAA).²⁰ Thus by the end of the BDM phase all subjects observed the same distributional information in the aggregate.

To measure the subjects' levels of risk aversion, the second phase implemented the Holt-Laury procedure presented in Holt and Laury (2002), with subjects selecting in each of ten rows whether they would prefer the "safe" (i.e., less variable) or "risky" lottery.²¹ To simplify the analysis of subjects' inferred levels of risk aversion, the exact same lotteries from the low-payoff treatment in Holt and Laury (2002) were used without modification. That is, the safe option had possible outcomes of \$2.00 and \$1.60, while the risky option had possible outcomes of \$3.85 and \$0.10. Similar to the BDM phase, subjects were told that one of the rows would be randomly chosen, with their chosen lottery in that row resolved and paid.

After submitting their responses for the Holt-Laury task, the resolution procedures were explained on screen (see Figure A.4 of Appendix 2.A). Although it would have been straightforward to automate the resolution of the various randomizations to determine the subjects' respective payoffs, to ensure the credibility of the randomizations each subject in turn—and in private—rolled seven unique dice simultaneously: one 20-sided die to determine which stage in the BDM phase was resolved, three 10-sided dice to yield a three-digit decimal number which determined the random price based on the relevant distribution for that stage, one 10-sided die to determine which row of the Holt-Laury procedure was played, and two 10-sided dice to determine the outcome of the lottery chosen by the subject in the randomly selected row of the Holt-Laury procedure.²²

Following the payoff resolution, the subjects completed a ten-question survey on

²⁰The six repeated distributions were #1, #2, #4, #5, #9, and #14.

²¹Figure A.3 of Appendix 2.A shows a screenshot of the implementation of this procedure.

²²While any of the lotteries in the Holt-Laury procedure could have been resolved with one ten-sided die, two ten-sided dice were used to make that part of the randomization more intuitive for the subjects given that the probabilities of the lotteries had been expressed in percentages.

SurveyMonkey, which included (among others) questions about their valuations for the item before and after the experiment, how they came to make their decisions, and the clarity of the instructions.²³

2.4 Descriptive Statistics and Results

Three sessions were conducted in the Xlab in February and March of 2011 with a total of 69 subjects. Each session lasted slightly less than one hour. The average monetary payment was \$15.86, and 22 (i.e., 31.9%) of the 69 subjects also received the item in addition to the monetary payment (see Table 2.2).

Since some of the subjects returned to previous rounds and revised their submissions, the average number of WTP observations per subject exceeded 20.²⁴ As seen in more detail in Table 2.2, 22 subjects (31.9%) made no revisions. Among the subjects who made any revisions, the mean number of revisions was 7.8 and the median was 5.²⁵

Over all subjects the mean of the last WTP submissions in each round was \$4.12, with a standard deviation of \$3.68. Elicitations ranged from a minimum of \$0 to a maximum of \$15—the entire endowment. The average price drawn over the 69 randomly realized BDM mechanisms was \$6.91. For the 22 subjects who obtained the item the average price was \$4.36.

The expected general patterns were observed in the Holt-Laury phase of the experiment: most subjects chose the safe option for the first row (with only a 10% chance of receiving the higher payoff from either lottery), chose the risky option for the tenth

²³See the end of Appendix 2.A for a full list of the survey questions.

²⁴Subjects were free to resubmit the same WTP in a given round. Perhaps frustrated by a slow-loading webpage, a few subjects were in the habit of entering the same amount for a given round several times in quick succession. These redundant observations were cleaned from the data set.

²⁵Nearly a third of the subjects never revised their earlier bids at all. Insofar as there was variation over the first bids, learning seemed to take place for many of the subjects. This raises the question of whether or not the process—or potential desirability—of making revisions was not sufficiently transparent or salient. If I were to repeat this experiment I would add a final screen to the BDM phase that would display all 20 distributions viewed by the subjects together with their latest submitted bids and allow them to make final revisions to any previous round before confirming their responses for the phase. Most models would predict that this modification would produce weakly more revisions per subject and bring last bids weakly closer to each other—within subjects, at least.

row (where the higher outcome of either lottery was certain), and crossed from safe to risky lotteries only once moving from the first to tenth rows (typically in the sixth or seventh row). The data for the Holt-Laury phase suggest that of the 69 subjects in this experiment, 13.04% were risk loving, 13.04% risk neutral, and 73.91% risk averse.

Table 2.3 presents the responses to the categorical survey questions by the type of instructions. The counts are quite similar across instruction type for nearly all responses, the outlier being that twice as many subjects with detailed instructions reporting that they felt “Hungry” (10 versus 5 out of 68).²⁶ A majority of the subjects reported that they liked cookies (51.5%), with nearly a quarter more of them reporting that they strongly liked cookies (23.5%). More subjects reported a downward trend in their valuations over the course of the experiment than an upward trend (19 versus 10), but the modal response was that their valuation remained about the same (32 out of the 68 respondents).²⁷

Descriptive statistics for the average *last* elicitations of WTP for each of the 14 distributions appear in Table 2.4.

2.4.1 Results on Dependence

Exactly three of the 69 subjects submitted a constant elicitation for all 20 rounds, consistent with the standard model with a certain valuation (see the charts for subjects 41, 52, and 56 in Appendix 2.B). Six more subjects had zero variance in their last elicitations for each round, meaning that at some point they went back to revise one or more submissions so that they would be perfectly consistent across rounds—consistent with the standard model with learning (see the charts for subjects 13, 18, 29, 44, 50, 64). Nevertheless, Figure 2.2.1 and the bulk of Appendix 2.B show that the vast majority of subjects exhibited positive variance in their elicitations to a greater or lesser degree, even over their last elicitations.

The visual suggestions of widespread distributional dependence in Appendix 2.B are backed up by regressions on the within-subjects data. Under the simple model

²⁶One subject failed to fill out the survey, so there is a maximum of 68 rather than 69 observations for the survey data.

²⁷Collectively, the individual patterns of the subjects’ reported valuations over the course of the 20 BDM rounds roughly match their respective reports on the trend of their valuations given in the survey.

$WTP = \beta_0 + \beta_1 \cdot Median + \varepsilon$ (where *Median* is the median of the distribution displayed in a given round) the coefficient β_1 should be zero in the absence of distributional dependence. Running this regression specification separately for each of the 69 subjects, β_1 is found to be significantly different from zero at a 5% level for 36 subjects and at a 1% level for 33 (see Table 2.5).²⁸ That is, roughly half of the subjects gave responses that were sensitive to the distributions of random prices they faced each round. summarizes the numbers of subjects for whom the coefficient β_1 is significantly different from zero by instruction type, and it also gives the results of the separate regressions run on only the distributions with high or low support and on only the first ten or last ten rounds.²⁹

One possible explanation for the data in Figure 2.2.1 and Appendix 2.B is that subjects may have randomly changed their responses across rounds due to boredom. The AA/AB design implemented in the first two rounds was implemented in part to investigate the potential for variation in responses merely due to the repeated nature of the task (regardless of distribution). In the first round all subjects faced distribution #14 (a triangular distribution with high support and more mass to the right). In the second round 34 of the 69 of the subjects faced this same distribution, while the remaining 35 faced distribution #5 (with the same high support but a pronounced spike at \$5).

Figure 2.2 presents histograms for the *first* subject elicitation (i.e., the unrevised initial submissions) for the first and second round of the BDM phase. Panels A and B show the results of the AA/AB design. Panel A shows that—as expected—the distributions of the responses are essentially the same during the first round for these two groups of subjects (as yet undifferentiated by distribution). This is confirmed by a Kolmogorov-Smirnov test ($p = 0.955$). In contrast, Panel B shows that the distributions of responses are noticeably different depending on which distribution is viewed in the second round, and this difference is significant (the relevant KS test has $p = 0.050$). Were subjects simply choosing randomly across rounds this difference would not be expected to be significant.

²⁸The sample for each of these regressions was restricted to the subjects' *last* submissions for each round.

²⁹Parallel regressions using *Mean* as the regressor yield virtually identical results.

Panel C of Figure 2.2 breaks down the first elicitation for the second round further by splitting these two groups into subgroups based on what type of instructions they viewed. The graph shows that those subjects who faced the spiked distribution during the second round clustered at and around \$5 in similar numbers regardless of whether they had viewed basic or detailed instructions. The other salient feature of Panel C is the fact that the subjects with detailed instructions reported elicitation of \$0 more often than those with basic instructions regardless of the distribution faced in round 2.

2.4.2 Results on Mass

The data provide compelling evidence that the position of mass in the distribution of random prices can have a pronounced effect on the WTP elicitation across subjects. More mass to the right often leads to higher WTP elicitation. This is suggested at first glance by the descriptive statistics given in Table 2.4. The pairwise comparisons that appear in Table 6, Table 2.7, and Table 2.8, often show significant difference in average WTP across distributions. This is true even though these tables are based only on the last elicitation, which should exhibit less distributional sensitivity than either the first elicitation or the entire set of observations. Furthermore, the patterns and conditions under which the null statement of Proposition 2 is rejected are enlightening.

Table 2.6 holds the support and the type (i.e., shape) of the distribution constant and focuses on symmetric shifts in the location of mass within the fixed support. For five of the six resulting pairs, the null of Proposition 2 is soundly rejected. Interestingly, the effect disappears in all cases for the subjects who viewed detailed instructions.

Table 2.7 holds the shape of the distribution and the location of the mass fixed and varies the support across the pairs considered. For all of the seven pairs of distributions the null of Proposition 2 is rejected outright. That the effect is particularly strong when the support is varied is unsurprising, since the elicitation for the distributions with higher support would mechanically be higher if enough of the subjects remained unsure about whether it would be possible or proper to submit values out-

side of a given distribution’s support. This table, then, provides the clearest evidence that the detailed instructions had a substantial effect: for that set of subjects only two of the seven pairs had significantly higher elicitations for the distribution with higher support, and the level of significance was less pronounced.

Table 2.8 holds the support and the approximate location of the mass fixed, while varying the shape of the distribution. For these pairs the degree to which one distribution in a pair stochastically dominates another is much smaller than for the pairs considered in Table 2.6 or Table 2.7. Accordingly, the differences in elicitations are smaller and less often significant. The null of Proposition 2 is rejected only three times over the 24 pairs considered for all subjects—and on two of these occasions only at a 10% level. Interestingly, the pairs which exhibited significant differences all had low support. This pattern also holds for the two additional pairs with significantly different elicitations among the subjects with basic instructions. Also, the same general pattern of less significant responses to changes in distribution holds for the subjects with detailed instructions, as seen in Table 2.6 or Table 2.7.

The results of the within-subjects regressions reported in Table 2.5 also speak directly to the directional effect of mass on subjects’ responses. The sign of β_1 gives the direction of the bias: $\beta_1 > 0$ indicates mass-seeking bias, and $\beta_1 < 0$ mass-fleeing bias. Of the 36 (33) subjects whose β_1 is significantly different from zero at a 5% (1%) level, 5 (4) of them had a negative β_1 coefficient. That is, only a small fraction of subjects exhibited mass-fleeing bias. In the overwhelming majority of cases of distributional dependence, Proposition 2 can be rejected. The data show that more mass to the right seems to pull elicitations to the right.

2.4.3 Results on Comprehension

The purpose of using two randomized versions of the instructions was to vary the level of subject comprehension about the BDM mechanism, with the more explicit, detailed version serving as an attempt to raise the average level of comprehension of those subjects who received that set of instructions. The results of the last survey question: “Were the instructions for Phase I clear?” give a quick check to see if the subjects themselves—ignorant of the existence of another form of instructions—happened to

rate the clarity of the instructions differently according to their instruction version. As seen in Table 2.3, there is no definite pattern in the uninformed self reports that would confirm an unambiguous success of the detailed instructions to produce an improvement in subject comprehension (or, at least, an improvement in the subjects' *perception* of their comprehension). A robust ordered logit testing whether viewing the detailed instructions increased the category of clarity of subject responses to this survey question yields a z -score of 0.85 with a corresponding p -value of 0.394.

The round-by-round BDM data yield compelling evidence that the instruction version affected the magnitude of their elicitations. As displayed in Figure 3, for 19 of the 20 BDM rounds (all except the 12th) the mean WTP reported for subjects who viewed the detailed instructions is lower than the corresponding mean for the subjects who viewed the basic instructions.³⁰ Since there is no obvious prediction about the direction of any effect of detailed instructions on the magnitude of the subject responses, Table 2.9 reports the results of two-tailed tests of means—looking at each round separately—for the data presented in Figure 3. These tests show the differences in means across instruction versions are statistically significant at the 10% level for 8 of the 20 rounds for the set of subjects' first elicitations and 9 of the 20 rounds for the set of subjects' last elicitations.

This pattern of reduced average elicitations for the subjects with detailed instructions is primarily due to the fact that the subjects with detailed instructions consistently gave elicitations of \$0 more often than their counterparts with basic instructions. Figure 2.4 plots the last elicitation for each subject in each round by the type of instructions, showing larger circles when multiple observations coincide. As can be seen, the range of elicitations remains approximately the same across the type of instructions. For the subjects with basic instructions the modal elicitation for is \$5 for 7 of the 20 rounds, and for 5 of the rounds \$0 and \$5 have the same number of observations among this set of subjects. In contrast, the modal elicitation for subjects with detailed instructions was uniquely \$0 for every one of the 20 rounds.

The evidence presented in Figure 2.3, Table 2.9, and Figure 2.4 convincingly show that the level of detail of the instructions clearly influenced the magnitudes of subject

³⁰Figure 2.3 compares the means of the last WTP observations from each round, but the general pattern holds (often stronger) for the first WTP observations from each round.

elicitations, contrary to the null formulation of Proposition 3a.

The level of detail of the instructions also seems to have an effect on the variance of subjects' reported WTP over the 20 BDM rounds. The mean of the variances of the WTP elicitations of the subjects with basic instructions is 7.25 (with a corresponding standard error of 1.18), and the mean of the variances of the WTP elicitations of the subjects with detailed instructions is 5.22 (with a standard error of 1.20). Figure 2.2.1 also shows that the variance of elicitations was generally lower for subjects who viewed the detailed instructions. A one-tailed test of means indicates that this pattern is nearly—but not quite—statistically significant: the variance of the detailed-instruction responses is determined to be higher with a probability of only 0.117.

2.4.4 Heterogeneity

Apart from illustrating the presence of distributional dependence among many of the subjects, the charts in Appendix 2.B and the regression results summarized in Table 2.5 strikingly emphasize the heterogeneity in subject responses. Some subjects gave very stable responses that varied little with the distributions, while others fluctuated wildly as each round presented a different distribution. Most of the subjects who exhibited distributional dependence exhibited mass-seeking bias, but the responses of a few were mass-fleeing.

Figure 2.5 captures this heterogeneity by plotting the correlation of reported WTP and the median of the distribution against the correlation of reported WTP and the mean of the distribution for each subject.³¹ There are three major groups. First, there are 33 subjects whose responses don't correlate strongly with the mean or median of the successive rounds, appearing as points near the center of the graph (from correlations of between -0.5 and 0.5).³² They are relatively insensitive to the distributional information. Second, there are 29 subjects who are very responsive to distributions in a mass-seeking way, represented by points to the upper right of the graph (with correlations above 0.5). Lastly, there are seven subjects who are quite

³¹This figure also illustrates how closely the respective correlations with mean and median are matched.

³²The nine subjects who report a constant valuation in their last elicitations are not represented in the graph, since their correlations are undefined. Qualitatively speaking, they clearly belong in this group.

responsive to distributions but in a mass-fleeing way, represented at points on the lower left part of the graph (with correlations below -0.5).

As also seen in Table 2.5, Figure 2.5 illustrates that the level of detail of instructions has noticeable effects. First, more subjects with basic instructions are strongly sensitive to the distributions of random prices. Second, most of the subjects who exhibit mass-fleeing bias viewed detailed instructions.

The survey data offer more evidence of heterogeneity as well as some insights on the effects of the instruction types. The seventh question of the survey asked subjects: “How did you decide what values to submit each round?” The subjects provided free-form responses of varying length and detail, which I categorized after the fact. The responses to this question by all subjects appear exactly as given in Appendix 2.C, and a summary of these responses by category and instruction type appears in Table 2.3.

The four points farthest to the lower left on Figure 2.5 correspond to the four subjects reported in Table 2.5 as having coefficients of β_1 significantly negative at the 1% level.³³ All of these subjects viewed detailed instructions. Three of them indicated on the survey that they used a bimodal strategy; that is, they placed a positive bid when the support was low, and then to ensure that they wouldn’t receive the item on rounds with high-support distributions they reduced their submissions, often to zero.

Six of the eight subjects who reported using a bimodal strategy viewed detailed instructions. It seems that more subjects with detailed instructions were aware *that* they could avoid undesirably high prices when faced with unfavorably high distributions of prices. However, the data are ambiguous as to whether or not these subjects were more aware of *how* to avoid paying an unfavorably high price. For example, subject 21 (with detailed instructions) would have had an equal, zero-percent chance of paying five or more dollars by submitting reports of \$2.50 (in line with his or her other reports) rather than the \$0 actually reported when faced with distributions of high support.

In stark contrast, nine of the eleven subjects who reported intentionally targeting the mean, median, other quartile, or spike of the distributions viewed basic instruc-

³³These four were subjects 8, 21, 33, and 68.

tions.³⁴ While it is not apparent from their responses why these subjects chose this kind of strategy, this is the height of distributional dependence. These subjects are admitting—and demonstrating through their submissions—that their reported WTP is completely divorced from the inherent valuation they might place on the item.

A plausible explanation for the bizarre strategy of targeting the mean, median, or other feature of the distributions is experimenter-demand effect. On observing that the distributions varied across rounds, these subjects may have made the conjecture that the “correct” response was to vary their bids in line with the nature of the distributions. This would also explain why more subjects with basic instructions reported following this sort of strategy; they were not as thoroughly trained to simply submit their sincere valuation regardless of the distribution.

As seen in Table 2.3, the remaining categories of explanations were offered in similar counts across the instruction types. These other explanations included—among others—reporting a relatively fixed (and possibly predetermined) amount, always bidding zero or other very low amount, chasing the satisfaction of winning, and bidding “randomly”.

2.4.5 Results on Testable Predictions

As seen in Section 2.4.1, only three subjects submitted bids consistent with them being rational expected-utility maximizers who have certain and unchanging valuations: that is, both their first and last bids were identical across all 20 rounds of the BDM phase (within subject).³⁵ Six more subjects had some variation in their first bids but no variation in their last bids for each round, consistent with the standard model with learning.³⁶ Collectively, the behavior of these nine subjects is also consistent with any of the models presented, provided that the magnitude of the purported individual biases is small in relation to the discreteness of the distributions used. Put another way, to a greater or lesser degree 60 of the 69 subjects submitted bids that did not conform with the predictions of the standard model.

³⁴Subjects 17 and 66 reported targeting the mean; 7, 19, 22, 31, 40, and 51 the median; 26 the first quartile; 67 the fourth quartile; and 2 and 35 the highest probability in the distribution. Of these, 17 and 40 viewed detailed instructions.

³⁵Again, see the charts for subjects 41, 52, and 56 in Appendix 2.B.

³⁶See the charts for subjects 13, 18, 29, 44, 50, 64 in Appendix 2.B.

Looking at the across-subjects data, there is no support for the prediction made by the model of Horowitz that elicitation would flee mass. As seen in Table 2.6, Table 2.7, and Table 2.8—which look only at last elicitation to help control for possible effects of learning—the distribution with more mass to the left *never* yields a higher average WTP that is in any way remotely significant.³⁷

The within-subjects data show relatively few of subjects exhibit mass-fleeing bias. Of the seven with correlations of less than -0.5 in Figure 2.5, not one of them clearly represents the case predicted by Horowitz and Kőszegi-Rabin, where a subject with a relatively high valuation would increase her reported WTP as a result of asymmetric regret or the attachment effect.³⁸ In contrast, one of the subjects specifically describes being motivated by the *comparison* effect: “I did not want to pay more than 5. If it was very likely to be less than five my willingness to pay went down because I expected it to be less than 5.” That is, the subject’s WTP decreased when prices were more likely to be low, since paying her initial WTP would feel like a loss compared to the low prices that would likely prevail.

Similarly, the across-subjects data do not support the predictions of Lusk, *et al.* As Table 2.6 shows, while it is true that the triangular distributions with low support (#11 and #12) produce the strongest significant effect on willingness to pay seen holding the support and distribution type constant, the effect has the *opposite* sign of that predicted by the model of Lusk, *et al.* Distribution #12 received last bids that were higher than those of distribution #11, and this effect was significant at a 1% level.

The primary prediction of Kaas and Ruprecht is that subjects uncertain of their valuations will generally bias their elicitation downward, fearing the specter of overpaying for an item more than forgoing the item and any attendant surplus. The survey conducted at the conclusion of the experiment asked subjects—retrospectively—what their dollar valuation was for the item was prior to the beginning of the BDM phase

³⁷The closest case would be where distribution #7 is compared to distribution #11 in Panel A of Table 2.8.

³⁸Of the seven, six report average WTPs that are lower than the mean or median of all of the distributions used. Subject 68 had the highest average WTP at 3.225, but as seen in Appendix 2.B this subject is giving purely bimodal responses that reflect the support of the distribution; there is no mass-fleeing response within the set of distributions with low support.

(*pre_wtp*) and (separately) what their dollar valuation was for the item prior to the payoff resolution (*post_wtp*). Comparing the mean WTP of each subject's last bids over all 20 BDM rounds to their respective responses to these survey questions gives some measure of their "underbidding" or "overbidding" during the BDM elicitations. Across all subjects, this method suggests that 60.61% of them (40 out of 66) were underbidding on average.³⁹ Curiously, this rate of underbidding is exactly the same regardless of instructions type: though the magnitudes of underbidding were different, the rates of underbidding for subjects facing basic or detailed instructions were identically 60.61% (20 out of 33 in each case).⁴⁰

The other main prediction of Kaas and Kuprecht is that subjects who are more risk averse will experience a greater downward bias in their bids. This is in direct contrast to the prediction that would be made if subjects were presumed to be treating the BDM mechanism as a first-price, sealed-bid auction, which would predict higher reported valuations with increased risk aversion other things being equal.⁴¹ Figure 2.6 displays the average WTP of each subject as a function of the number of safe choices made in the Holt-Laury procedure, which serves as a proxy measure of risk aversion. The figure shows that there is a negative relationship between risk aversion and elicited WTP, as predicted by the model of Kaas and Ruprecht. As seen in the superimposed trendlines in Figure 2.6, this negative relationship holds more strongly for subjects who viewed basic instructions. However, simple OLS regressions show that the negative slopes of these trendlines are not significantly different from zero.⁴²

As seen in Table 2.5 and Figure 2.5, the data broadly support the implications of the first and second principles of the framework presented for a large subset of the subjects. Across subjects, more mass to the right for a given distribution of random prices typically implies higher average WTP reported in that round.

For the set of subjects described in Section 4.4 as responsive to the distributions of random prices, the third principle of the framework also holds. This can be seen

³⁹These calculations of underbidding use *pre_wtp*.

⁴⁰The total of these numbers is 66 rather than 69 since one subject failed to fill out the survey and two others submitted non-numeric data for the *pre_wtp* question.

⁴¹See Vickrey (1961) for a discussion of risk attitudes and equilibrium bidding strategies in first-price auctions.

⁴²Specifically, $p = 0.129$ for the subjects with basic instructions, $p = 0.317$ for the subjects with detailed instructions, and $p = 0.141$ for the regression with all subjects.

by plotting the differences between subjects' last reported WTP for each round and the average of their last WTP reports against the associated differences between the mean of the distribution and the average of their last WTP reports.⁴³ Figure 2.7 gives this scatterplot and the fitted Lowess curve for the 29 subjects with mass-seeking bias over each of the 20 rounds, and Figure 2.8 gives the same for the seven subjects with mass-fleeing bias. Constant effects of range or mass regardless of distance from a subject's valuation would imply that these Lowess curves would have constant slope (positive for subjects with mass-seeking bias and negative for those with mass-fleeing bias). Each of the fitted Lowess curves strikingly shows diminishing effects in either direction.⁴⁴

2.5 Discussion

As seen in Sections 2.4.1. and 2.4.2., this experiment provides substantial evidence for distributional dependence in the BDM mechanism. The existence and persistence of distributional dependence broadly calls into question how the BDM mechanism is applied and interpreted. One might reasonably wonder whether there might be a way to mitigate the effects of distributional dependence, such that subject responses are more stable and reliable. For instance, can this be done through increased subject comprehension, learning, or training? Should the distributional information simply be hidden from subjects?

2.5.1 Mitigation Through Comprehension

As seen in Sections 2.4.3. and 2.4.4, the level of detail of the subject instructions affects the magnitude and the variance of subject responses as well as—at times—the strategy used by the subjects. The significance of these results is surprising given how little the instructions actually differed. The detailed instructions only added a

⁴³Here the average of the last WTP reports is serving as a proxy for the sincere WTP, which would be the ideal.

⁴⁴For the seven subjects with mass-fleeing bias charted on Figure 2.8, the point of inflection occurs around 5 rather than 0 due to the very low average *last* WTP reports of these subjects and the bimodal strategy that many of them used.

diagram, a paragraph with a worked-through example, and an additional sentence on how to respond when one's valuation falls outside the support of a distribution.⁴⁵

These results present new questions: Why did the more detailed instructions systematically prompt lower overall responses? Is comprehension really the mediating variable, or is there some other channel through which varying the instructions affects the magnitude and variance of the WTP elicitation? More importantly, while the responses of the subjects with detailed instructions were more stable overall (relative to those of the subjects with basic instructions), do they more closely approximate the subjects' true valuation for the item on offer?

The scatter plot in Figure 2.4 clearly shows that rather than a general reduction of \$1 to \$2 in reported WTP for subjects with detailed instructions—as the means reported in Figure 2.3 might suggest at first glance—the average WTP differs significantly by the type of instructions due to consistently larger numbers of subjects with detailed instructions submitting a WTP of \$0.

One possible explanation for this disparity would be if significantly more subjects with detailed instructions truly began the experiment with valuations of \$0 for the item. Indeed, at first glance the survey data collected at the end of the experiment suggest that this might have been the case. All four of the subjects who stated that they disliked cookies to some degree or another viewed detailed instructions (see Table 2.3). If these responses weren't affected by the BDM phase, the Holt-Laury phase, or the payoff determination, this confluence of all four such subjects in the detailed treatment might be expected to happen randomly with a probability of about 6.6%.⁴⁶ However, it turns out that the pattern of increased \$0 submissions among subjects with detailed instructions cannot be largely attributed to the four subjects who disliked cookies. Out the 80 last bids of these four subjects over the 20 rounds of the BDM phase, only 22 were for \$0. The median of these 80 submissions was \$1, and fully a quarter of them were actually greater than \$4. Curiously, when asked retrospectively in the same survey about their willingness to pay for the gift certificate for a dozen cookies at the beginning of the experiment, only one of these subjects who reportedly didn't like cookies responded \$0. One said \$3 and the other

⁴⁵Again, see the boxed areas on the second page of the instructions appearing in Appendix 2.A.

⁴⁶That is, $(35/69)^4$.

two \$5. Perhaps they were thinking of reselling the certificate or sharing cookies with others.

Another explanation for this disparity is that the detailed instructions motivate some subjects to place a higher value on the endowment of \$15. As seen in the detailed instructions that appear in Appendix 2.A, the fields of the diagram of the illustrative diagram express the possible payments as “\$15” or “\$15 – X ” in a way that’s more salient than in the basic instructions. While the existence of the endowment and its relation to the final payment are mentioned in the basic instructions, the specific value of \$15 does not appear in the basic instructions. It’s possible that highlighting the numerical value of the endowment in the detailed instructions in this way may have led some of the subjects viewing them to place greater emphasis on obtaining the maximum monetary compensation—regardless of their preference or latent willingness to pay for cookies.

It may be that subjects who viewed the more detailed instructions were able to better grasp the complexities of the institution while simultaneously learning how to avoid those complexities and any attendant costs—computational or otherwise—that might arise from dealing with these complexities. That is, the more informed subjects may understand enough about the institution to know how to opt out of it, and given that knowledge more of them do so. Given the survey responses of the subjects, it is not clear that this increase in the number of bids of \$0 for subjects with detailed instructions more accurately reflects their true willingness to pay. Five subjects cited a desire to bid zero for the item as they explained their decision-making process in survey question #7, and of these two received detailed instructions.

While the detailed instructions yield more stable bids over the course of 20 rounds, this effect is largely driven by subjects who sidestep the BDM mechanism entirely by submitting bids of \$0, which the survey responses suggest do not always reflect their true valuations. More generally, this raises serious concerns about whether increasing subject comprehension of the BDM mechanism through better instructions would yield more reliable measures of subjects’ true WTP. Indeed, an increased proportion of opt outs by more informed subjects would likely lead to *worse* estimates of subjects’ sincere valuations.

2.5.2 Mitigation Through Learning

One potential cause of instability of subject responses is if they learn more about their valuation of the over the course of the experiment through the distributional information presented. This kind of learning would likely manifest itself in two ways. First, the variance of subjects' later responses would be weakly smaller than the variance of their earlier responses. Second, subjects would have less reason to revise their submissions in later rounds, and thus might make fewer revisions in later rounds.

A simple comparison shows that the variance does in fact diminish in the subjects' later responses. Figure 2.9 gives a scatter plot which compares the variance of the first elicitation over the first ten rounds with the variance of the first elicitation over the last ten rounds for all 69 subjects. It shows that the majority of the subjects had smaller variances over the last ten rounds. The mean of the 69 variances of the first elicitation for the first 10 rounds is 4.823, and the mean for the 69 variances of the first elicitation for the last 10 rounds is 6.592. The corresponding one-tailed t-test for these two sets gives a p -value of 0.0061, so the variances for the last 10 rounds are indeed significantly lower than those for the first 10 rounds.

Figure 2.2.10 shows the total number of revisions made by all subjects for each round. Although the number of revisions per never approaches zero, there are clearly fewer revisions made in later rounds. While there are 49 revisions in the first round, the number diminishes to 11 by round 9, after which the count remains at 11 ± 3 for the remaining rounds.

Each of these pieces of evidence suggest that subjects take into account past distributions when submitting a WTP report in a given round. But how important are past rounds? One possibility is that a priming effect could persist—that subjects might overweight the distributions in the early rounds throughout the remainder of the BDM phase.

The AA/AB set-up of the first two rounds allows for a weak test of this kind of priming: specifically, whether or not the consecutive appearance distribution #14 for 34 of the 69 subjects would leave a stronger impression that the value of the item was high relative to the inferences of the 35 subjects who viewed distribution #5 in the second round. Figure 2.2.11 shows that this is clearly not the case: as soon as

round 3 the subjects who had viewed the higher distribution #14 twice consecutively in rounds 1 and 2 have Looking at the first elicitation from each round from 3 to 20 separately, the sign of the t-test comparing the WTP elicitation across subjects who viewed distribution #14 twice consecutively with the responses of those who only viewed it during the first round is inconsistent and the magnitude is usually quite far from significant. In particular, round 3—which might be expected to exhibit the highest degree of priming—has a t-test statistic of 0.6495 (with a corresponding p-value of 0.2591), which is nowhere near significant and has the opposite-from-expected sign. There is no evidence for a weak priming effect lasting even a single round—much less over the duration of Phase I.

To the contrary, this study provides evidence that subjects underweight the distributional information of previous rounds and place undo weight on the distribution of random prices in whatever round they happen to be in. By the end of the 20 BDM rounds it might be expected that subjects will have learned nearly everything they can from the overall distributional information presented, and that any additional distributional information would be small and marginal. This suggests that if learning is the primary driver of distributional dependence, elicitation for rounds 19 and 20 should be very close. In fact, only 25 of the 69 subjects submitted the same value for over rounds 19 and 20, 12 of which were repeated valuations of zero. The remaining 44 (63.8%) of the 69 subjects reported different valuations across these rounds. The average absolute difference between the elicitation in round 19 and round 20 for these 44 subjects was 2.73 for subjects with basic instructions, 2.35 for subjects with detailed instructions, and 2.56 for the combined group of all 44. Given that the average elicitation among these 44 subjects was 4.30 in round 19 and 4.42 in round 20, these are very significant differences.

One interpretation of this result is the conclusion that learning does not eliminate distributional dependence for most subjects. Another interpretation is that these elicitation might be rationalized by a learning model that places an unexpectedly large weight on the current round's information. Learning—albeit via an unconventional learning model—might then be considered a potential source of distributional dependence for a subset of subjects rather than a factor that would help mitigate it.

Whether or not learning from distributional information during an experiment

diminishes distributional dependence, displaying multiple distributions to subjects (either simultaneously or sequentially) does not ultimately eliminate the issue of distributional dependence. Rather, it just moves the problem up a level: subjects' responses would perhaps be influenced by the *set* of distributions chosen by the experimenter.

2.5.3 Mitigation Through Training

Although every subject in this study faced multiple BDM elicitations, each experienced only one BDM resolution, which came after all BDM elicitations had been irrevocably made. This study does not address learning that may occur as subjects gain more first-hand familiarity with the possible outcomes of the institution. It is possible that subjects would alter their strategies in subsequent encounters with the BDM mechanism in the wake of experiencing regret from either paying too much for an item or not obtaining an item at an attractive price. It might be expected that subjects trained through multiple exposures to BDM resolutions with real stakes would submit more accurate and more stable elicitations, in line with evidence that market experience often reduces deviations from theoretical predictions in experimental settings.⁴⁷

Training might improve the accuracy and stability of elicitations through the channel of comprehension alone. Observing the institution in action firsthand might clear up any remaining confusion subjects have about how the price that they pay will be determined or whether or not they will have the option of renegotiating the sale or price later. Intuitively, training might be even more effective if subjects experience forgone gains or unnecessary losses in early encounters with BDM resolutions. That is, regret might be a more powerful channel for inducing more refined future elicitations than increased comprehension alone. This notion is reflected in the BDM literature, which often presumes that agents are driven to avoid or minimize *ex post* regret.

The survey data in this study do not offer a strong case for the presence of regret. Five subjects failed to obtain the item when the random price was lower than their average reported valuation over the 20 BDM rounds. Four obtained the item

⁴⁷See List (2003).

at a randomly drawn price that was higher than their average reported valuation over the 20 BDM rounds. Though these subjects' actual valuations for the item remain unobserved, these cases suggest that several subjects may have been in a position to experience *ex post* regret—at least theoretically. Nevertheless, out of the 68 subjects who submitted responses to the survey at the end of the experiment, only four reported feeling regretful (and only one of these was among the nine described above).⁴⁸ Moreover, three of the four subjects who reported being regretful simultaneously reported being pleased, and one of them actually obtained the item for a price of \$0—hardly a regrettable outcome (unless, perhaps, one is trying to avoid the temptation of a dozen freshly baked cookies).

Further study into the effects of training on BDM responses is warranted, and there are a number of unanswered questions. How many rounds of training are required? If training is effective, is regret the operative channel? Would training with a particular good cross over and improve subject responses in a similar way for a different item?

2.5.4 Mitigation Through Nondisclosure

If the distribution of random prices has the potential to influence subject responses in significant but unobserved ways, why not simply keep the distribution unknown to subjects at the time of elicitation?

From an empirical standpoint, the chief risk of not disclosing the distribution of random prices to subjects facing the BDM mechanism is a loss of experimental control. Insofar as a distribution of random prices that is known to the subjects might influence their responses, it may be possible to anticipate and account for such influence when interpreting any results. With a hidden distribution there is no telling what subjects might infer or imagine about it, resulting in reports of sincere, unobserved valuations that may have been influenced in an unknown direction and magnitude by an unobserved distribution of random prices perceived by subjects individually. In an experimental setting there is a nontrivial risk that some seemingly innocuous number unexpectedly serves as an anchor point for subjects' perceptions

⁴⁸Specifically, those who reported being regretful were subjects 19 (obtained the item at a price of \$5.50), 40 (obtained the item at a price of \$7.50), 49 (obtained the item at a price of \$0), and 66 (did not obtain the item at a price of \$13).

of the bounds or shape of the distribution.⁴⁹

Another—although probably lesser—concern is the credibility of the experiment and experimenter. Revealing the distribution of random prices to subjects offers greater transparency; hiding it has the potential to raise suspicion among subjects. Regardless of what their valuations for an item or their beliefs about probabilities might be, it's important that subjects in economics experiments can trust that they aren't being deceived. Offering up front to reveal the distribution after the resolution of the BDM mechanism or at the conclusion of the experiment (as in Bohm, *et al.*) seems an eminently appropriate precaution.

It is an open question whether nondisclosure will yield subject responses that are more accurate reflections of subject valuations, and further study should be undertaken.

2.5.5 Implications for Research

The findings of this experiment suggest that in designing experiments using the BDM mechanism, care should be taken not to use a distribution where the subject responses are expected to lie at one of the extremes of the support. Given the mass-seeking tendency of most subjects, when elicitations are found to be generally lower (higher) than the mean of the distribution of random prices used, added caution is warranted when drawing an inference that these responses are significantly higher (lower) than a given level.

As pointed out by Mazar, *et al.*, studies that rely on the BDM mechanism to obtain absolute estimates of valuations for specific goods will be the most sensitive to any effects of distributional dependence. Studies that use the BDM mechanism to determine merely *relative* valuations between goods will be more robust—provided that the same distribution of random prices is used for each of the goods being compared.

If the range or other attributes of a distribution are to be kept from subjects, the experimenters must carefully decide how to describe the BDM institution in way that is comprehensible but not deceptive. The results may be quite sensitive to any

⁴⁹For more on anchoring, see Tversky and Kahneman (1974).

broad generalizations that the experimenters might make about the distribution of the random prices—either while presenting the institution or while responding to inquiries from subjects on the matter. A concerted effort should be made not to offer anchoring points inadvertently during earlier portions of the experiment. Also, for the sake of credibility it would be advisable to offer to reveal the hidden information at a later time.

The second-price auction also offers an incentive-compatible method for eliciting valuations from subjects. If there is worry that the BDM mechanism is sensitive to the distribution of random prices used, it is only natural to wonder whether the second-price auction would be a better alternative. While this study does not directly address this important question, there is cause for concern. It is reasonable to suspect whether subjects' bids in a second-price auction might be biased in response to their beliefs about the distribution of random prices they face, arising from the perceived distribution(s) of their opponents' valuations as well as their opponents' perceived bidding strategy (which in turn may be biased in a similar fashion). In practice, mass-seeking bias may also affect responses to a second-price auction, and this should be investigated in light of this chapter's findings.

2.6 Conclusion

This chapter decisively demonstrates that the distributions of random prices used in BDM elicitation for privately valued, concrete goods matter. Changing the support and the arrangement of mass of the distribution from which prices are randomly drawn can profoundly and significantly affect the magnitude of subject responses—contrary to the predictions of the standard model—even when the effects of rational learning from the distributional information are taken into account.

Furthermore, the patterns of observed distributional dependence are found to be largely inconsistent with the predictions of the models of Horowitz and Lusk, et al. There is suggestive evidence from the survey that students completed at the conclusion of the experiment that supports the more generalized pattern of underbidding predicted by Kaas and Ruprecht. Collectively speaking, the measures of risk aversion indicate that subjects are not treating the BDM mechanism as a first-price auction.

The proposed mass-seeking framework provides a substantially better description of the empirical patterns among the subjects sensitive to distributional dependence, including the fact that these effects diminish with distance.

The level of detail of the instructions matters to a surprising degree. While the more detailed instructions yield responses that are more stable and less sensitive to distributional dependence, it is not the case that they necessarily more accurately reflect the true valuations of the subjects. Insofar as distributional dependence may pose complications or difficulties for eliciting valuations from subjects, the solution is arguably not to be found in simply increasing the level of detail of the instructions.

The evidence for distributional dependence and—especially—the demonstrated possibility that more detailed instructions might result in responses that are *farther* away from a subject’s “true” valuation raise more fundamental questions about the nature of valuation. Analogous to Heisenberg’s uncertainty principle from quantum physics, the act of measurement may have a nontrivial effect on the quantity being measured.⁵⁰ In this context, the more precisely we attempt to measure a subject’s valuation for a good the more we may in fact be altering that selfsame valuation. There is already an ample literature describing the resulting world of unstable and mutable valuations and preferences, and these results reinforce that view.⁵¹

In the final analysis, the BDM mechanism is and remains an important tool for measuring the valuations of subjects both inside and outside of the lab. This work aims to contribute to our understanding of how this oft-used and relied-upon methodology should be implemented and interpreted with more care and consideration.

⁵⁰See Heisenberg (1927).

⁵¹See, for instance, Ariely, Loewenstein and Prelec (2003), which explores “coherent arbitrariness” and the influence that irrelevant anchors may have on reported valuations.

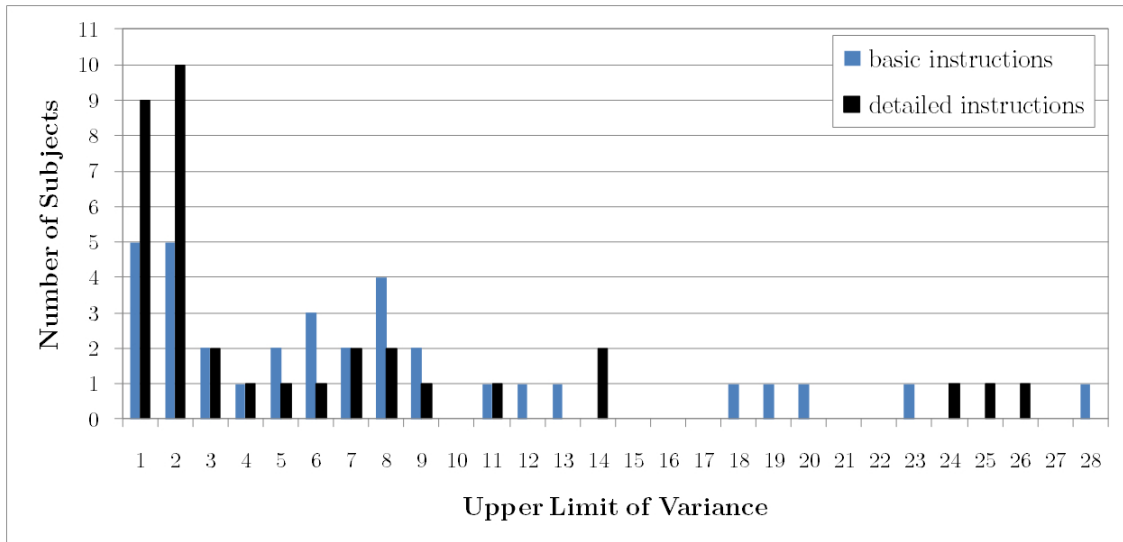
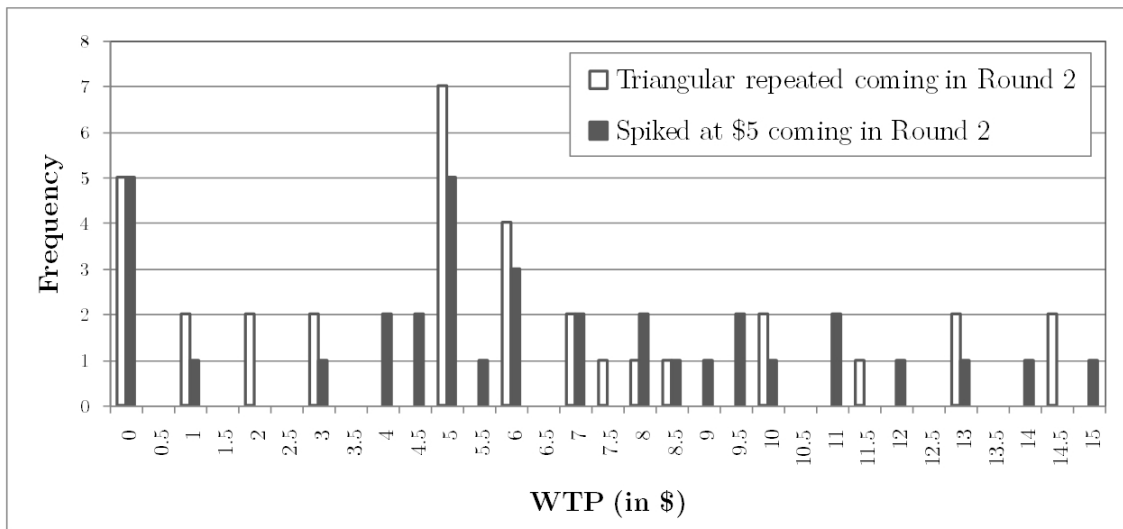


Figure 2.1. Paired Histogram of Variance of WTP Elicitations by Instruction Version

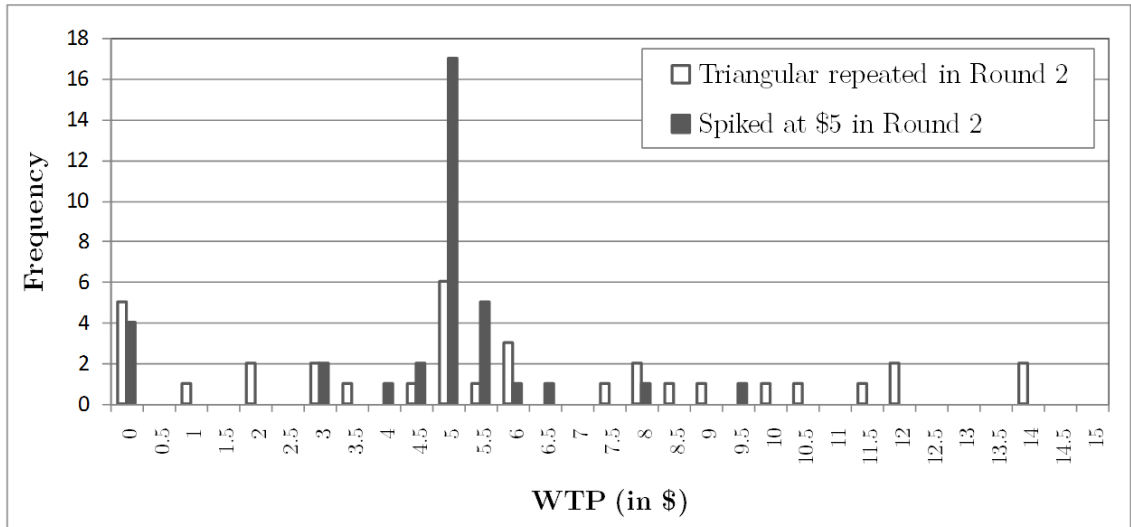
Figure 2.2. Willingness to Pay in Rounds 1 and 2

Panel A. Round 1 WTP (All subjects faced the same triangular distribution, #14.)

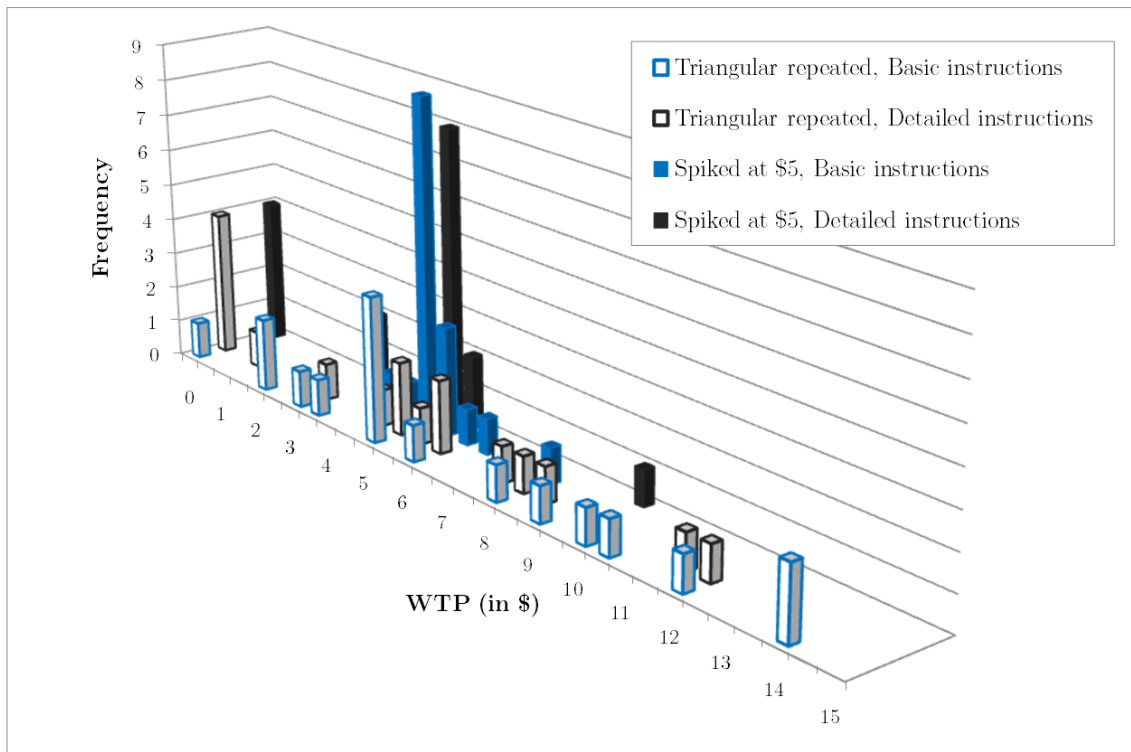


Chapter 2. Testing Distributional Dependence in the Becker-DeGroot-Marschak Mechanism

Panel B. Round 2 WTP (Half of the subjects faced a new distribution with a spike at \$5, #5.)



Panel C. Willingness to Pay in Round 2, by Level of Detail of Instructions



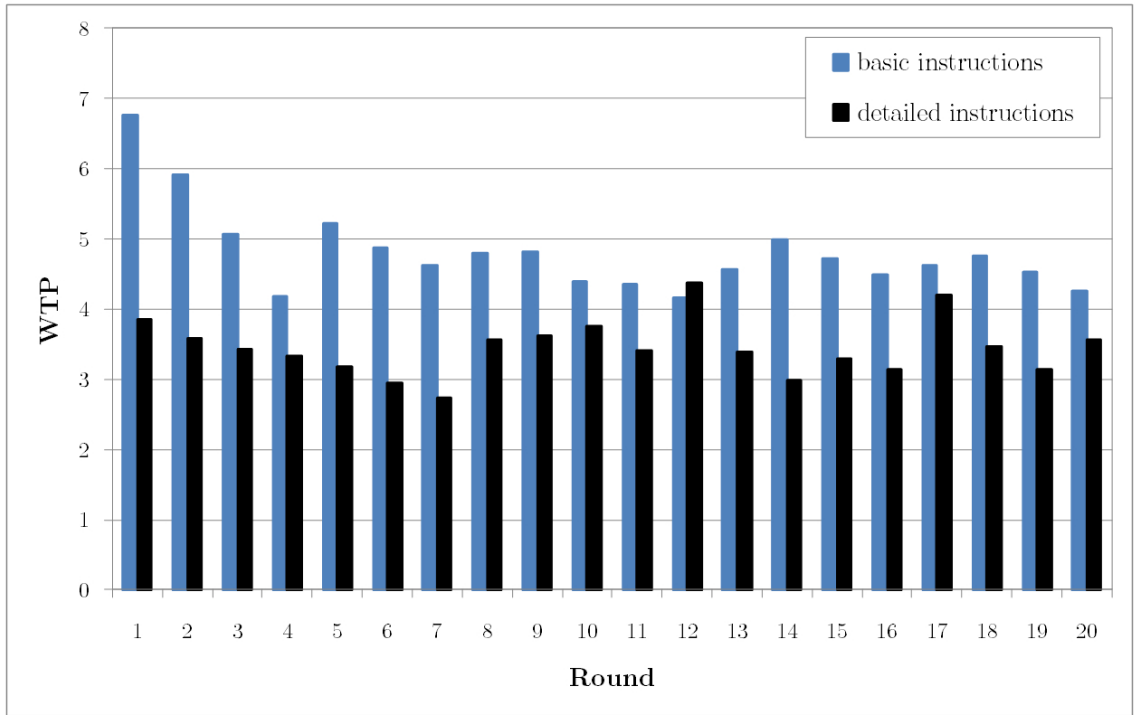


Figure 2.3. Means of Last WTP Elicitations in Each BDM Round, By Type of Instructions

Chapter 2. Testing Distributional Dependence in the Becker-DeGroot-Marschak Mechanism

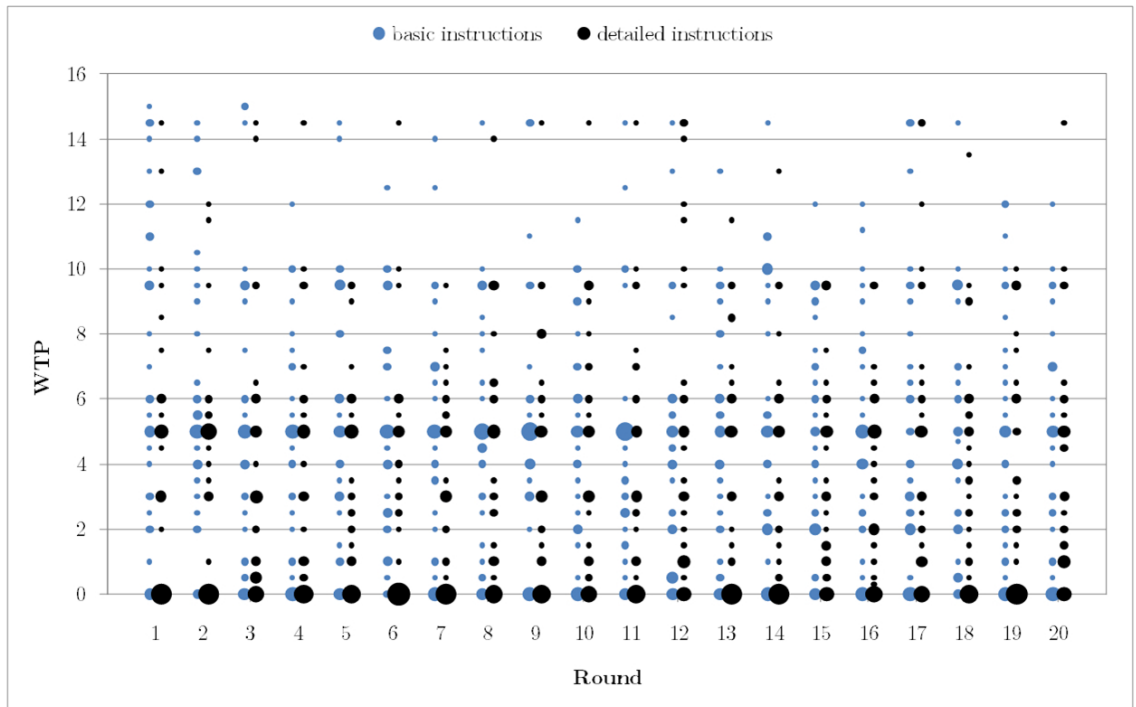


Figure 2.4. WTP Elicitations by Round and Instruction Version

Note: These are the *last* elicitations for each subject only. The number of observations at a given round and WTP level is proportional to the area of the circles (not the width). For example, in round 1 there were 4 subjects with basic instructions and 13 subjects with detailed instructions who submitted \$0.

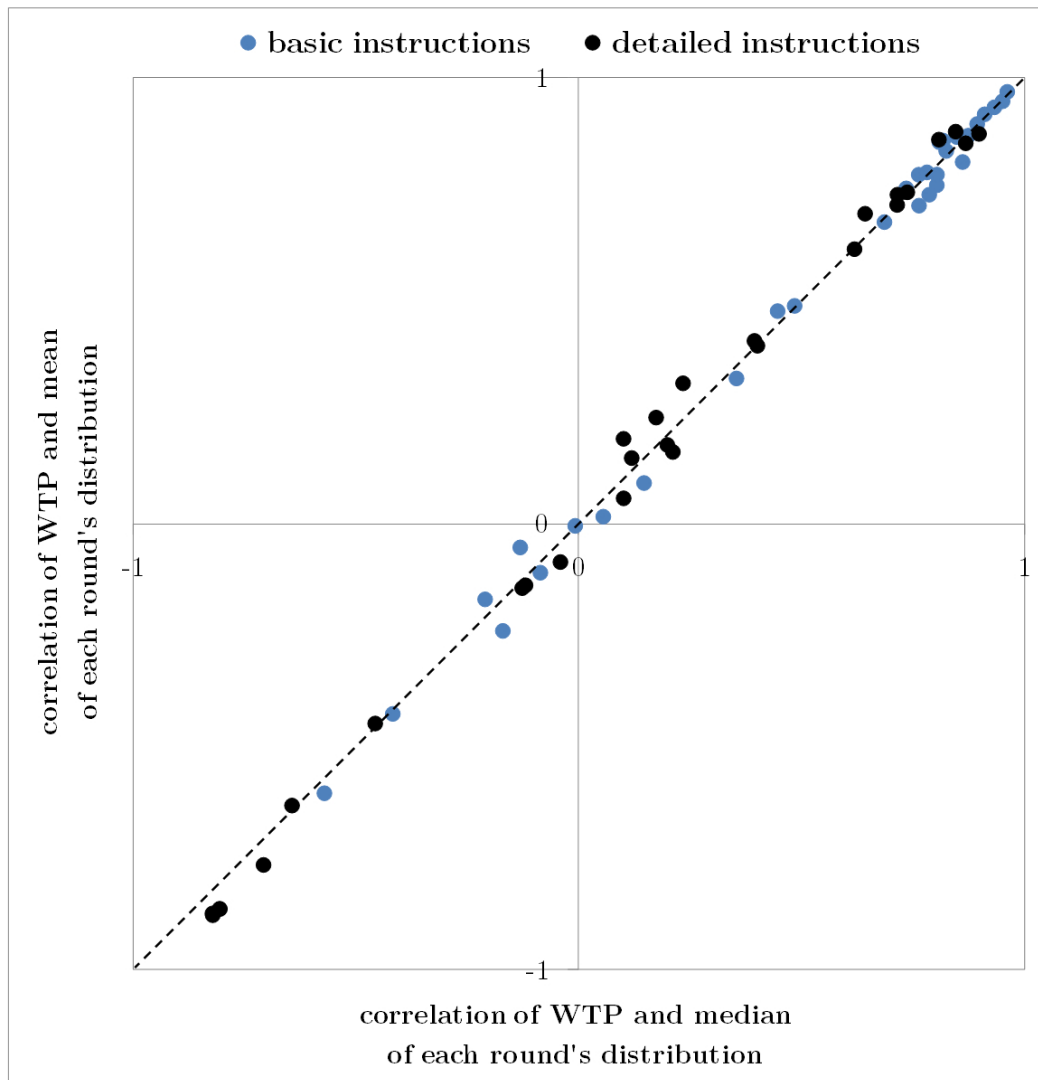


Figure 2.5. Correlation of Subject Reports of WTP to Median and Mean of Distributions, by Type of Instructions

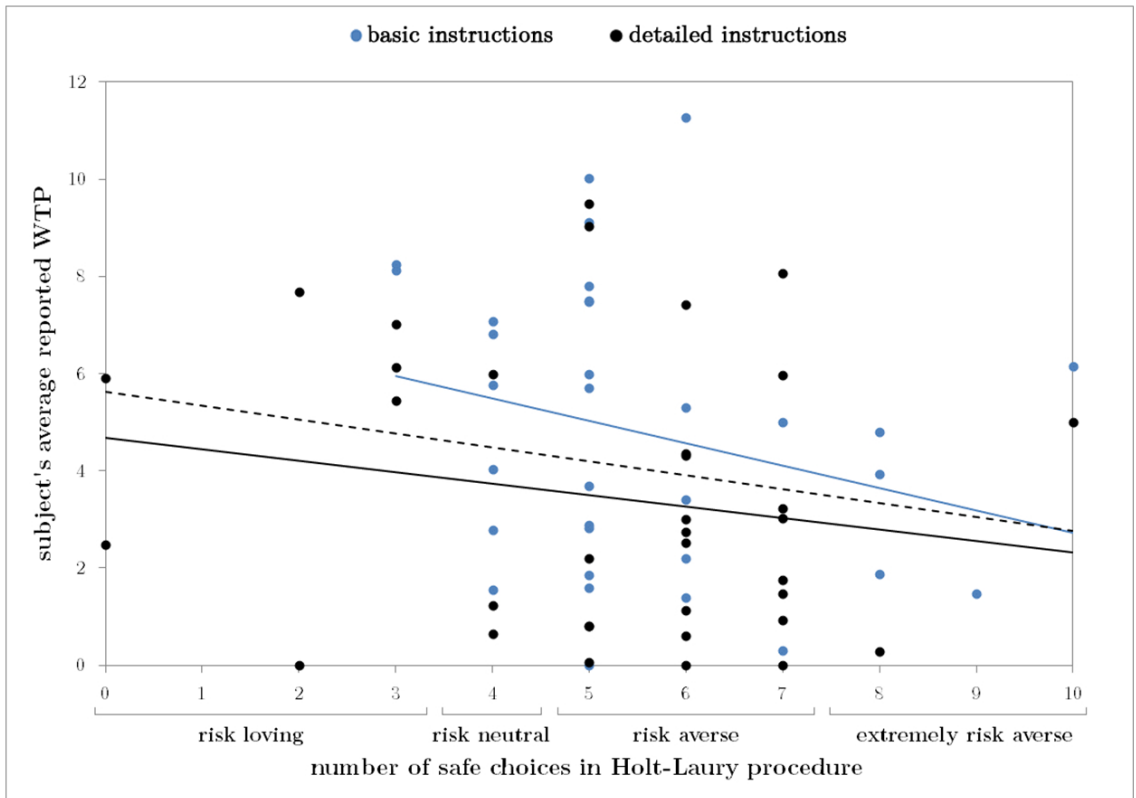


Figure 2.6. Measured Risk Attitude and Average WTP by Instruction Version

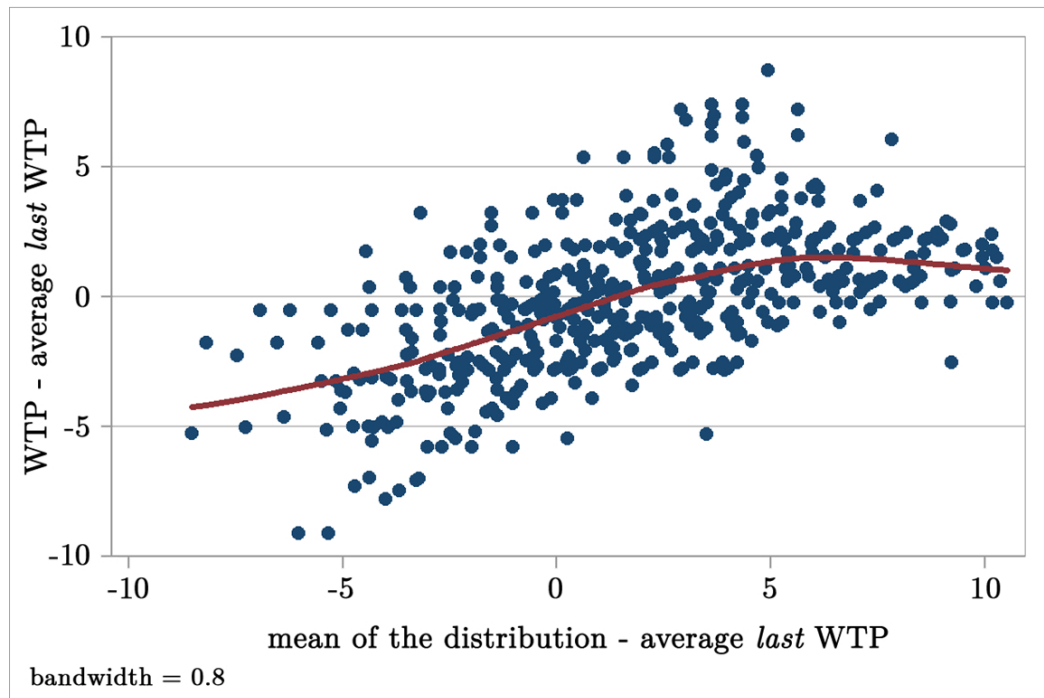


Figure 2.7. Fitted Lowess Curve for Responsive Subjects with Mass-Seeking Bias

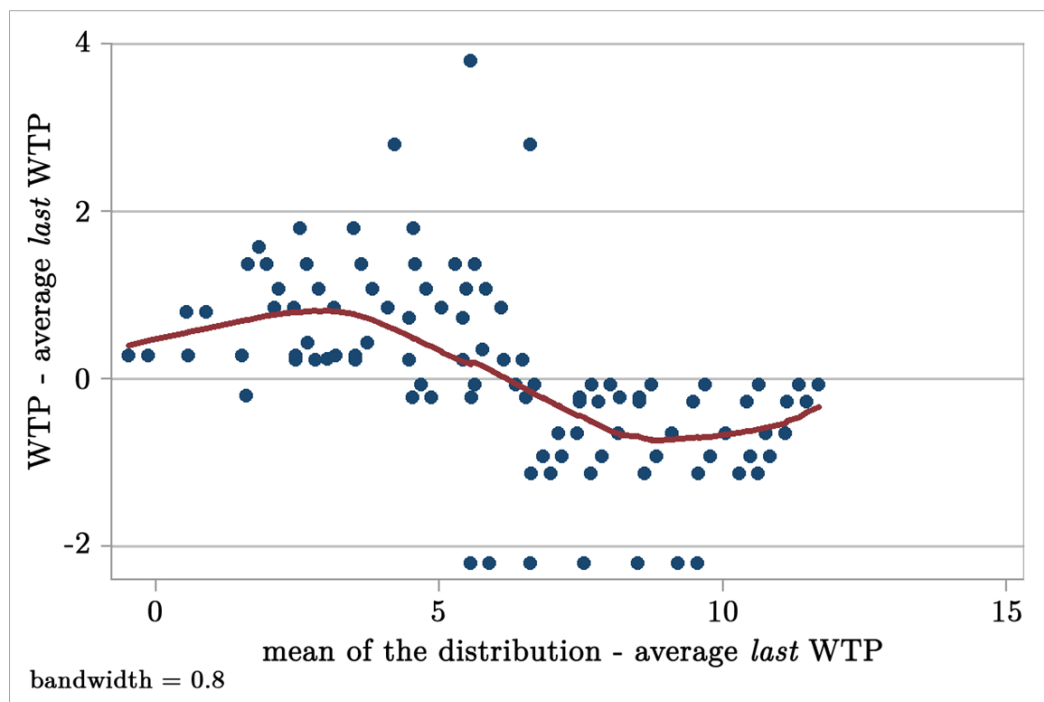


Figure 2.8. Fitted Lowess Curve for Responsive Subjects with Mass-Fleeing Bias

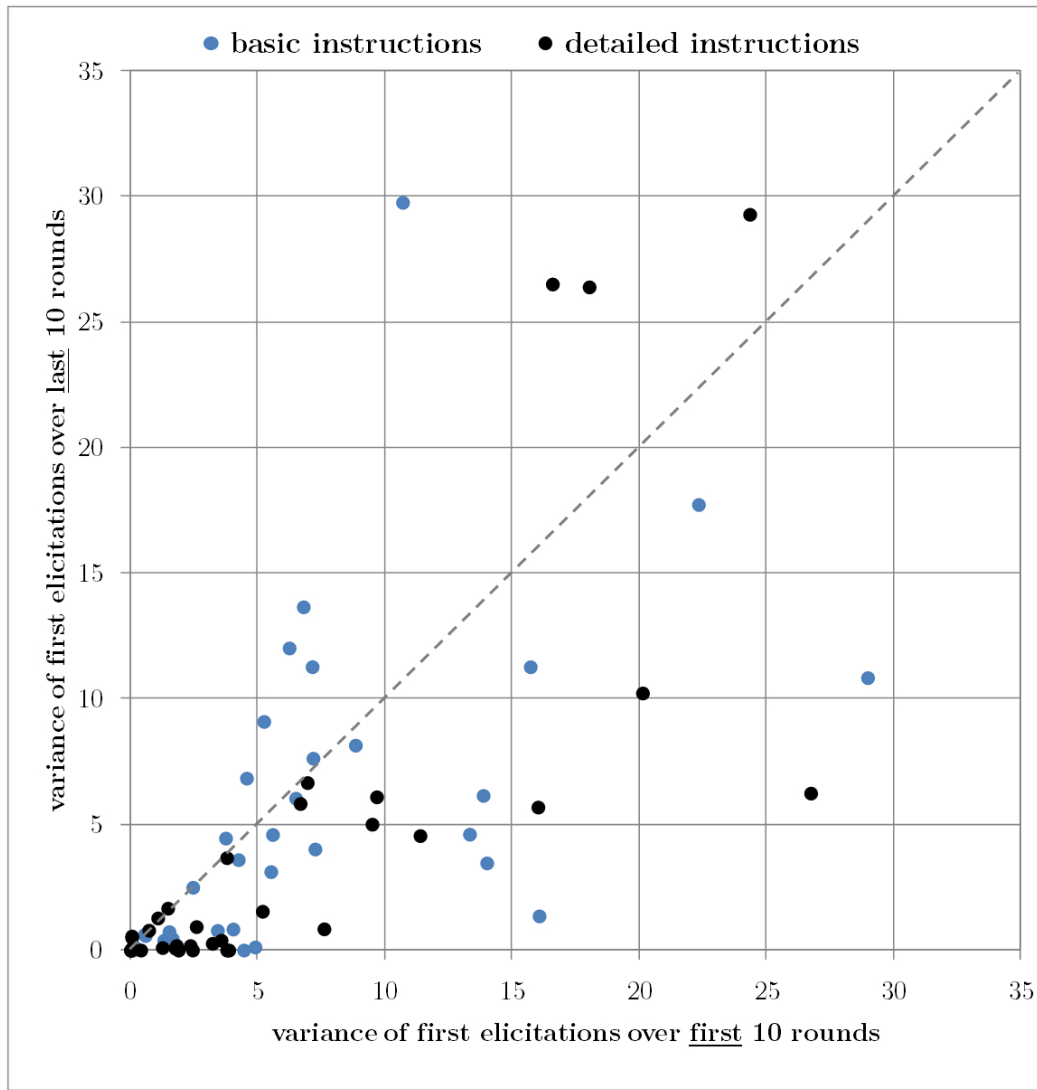


Figure 2.9. Scatter Plot of Subjects' Variances of BDM Elicitations: First 10 Rounds versus Last 10 Rounds, by Type of Instructions

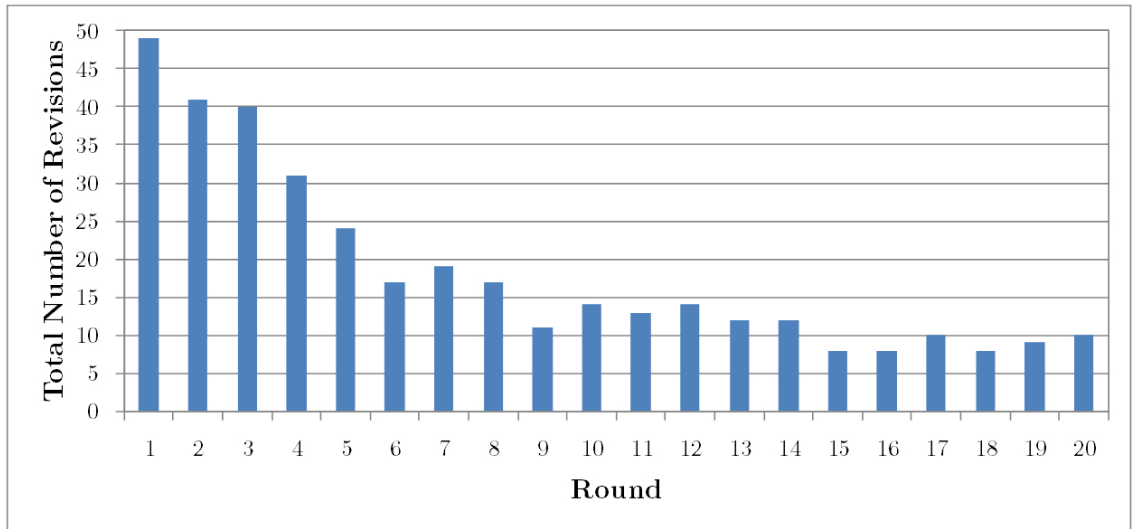


Figure 2.10. Total Revisions Made by All Subjects, by Round

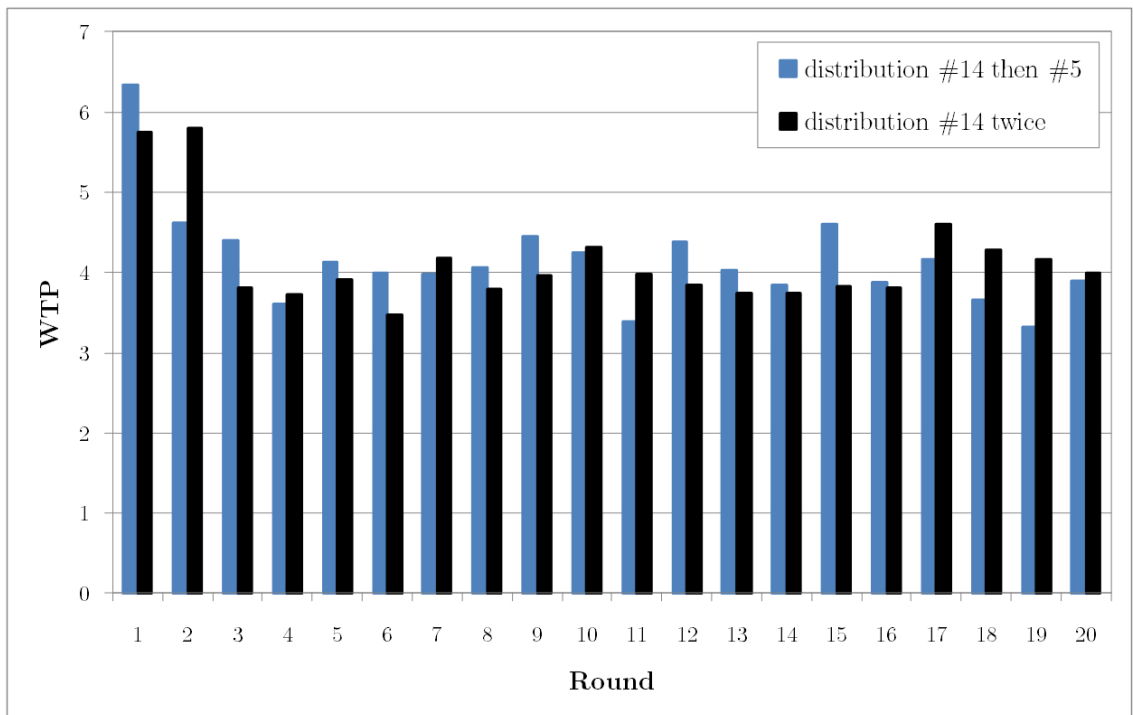


Figure 2.11. Mean WTP by Whether Distribution #14 Viewed Twice Consecutively in the First Two Rounds

Table 2.1. Summary of the Distributions Used in the Experiment

#	probability mass points (in thousandths)																Mean	Median
1	50	50	50	50	50	50	50	50	50	50	50	50	50	50	50	50	\$4.75	\$4.75
2																	\$9.75	\$9.75
3	240	40	40	40	40	40	40	40	40	40	40	40	40	40	40	40	\$3.80	\$3.50
4	40	40	40	40	40	40	40	40	40	40	40	40	40	40	40	40	\$5.70	\$6.00
5																	\$8.80	\$8.50
6																	\$10.70	\$11.00
7	90	90	90	90	90	90	90	90	90	90	90	90	90	90	90	90	\$2.75	\$2.50
8	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	\$6.75	\$7.00
9																	\$7.75	\$7.50
10																	\$11.75	\$12.00
11	97	93	87	83	77	73	67	63	57	53	47	43	37	33	27	23	\$3.09	\$2.50
12	3	7	13	17	23	27	33	37	43	47	53	57	63	67	73	77	\$6.41	\$7.00
13																	\$8.09	\$7.50
14																	\$11.41	\$12.00

Note: Blank spaces in the table represent points that occur with zero probability for the relevant distribution. Distributions #1, #2, #4, #5, #9, and #14 were each repeated once at some point over the course of the 20 BDM rounds for each subject.

Table 2.2. Summary Statistics by Session

	subjects	bids	revised	wtp	safe	item	price	payment	pre_wtp	post_wtp
session 1	20	527	16	4.43 (3.60)	5.35 (1.46)	8	7.10 (3.95)	15.37 (3.22)	6.65 (4.14)	4.97 (2.67)
session 2	24	602	14	3.88 (3.88)	4.75 (2.03)	7	6.64 (3.90)	15.56 (3.44)	6.80 (6.46)	5.61 (5.62)
session 3	25	618	17	4.09 (3.43)	5.96 (1.95)	7	7.02 (4.37)	16.53 (2.14)	4.57 (3.07)	4.47 (2.90)
all	69	1747	47	4.12 (3.68)	5.38 (1.86)	22	6.91 (4.04)	15.86 (2.97)	5.91 (4.74)	4.97 (3.85)

Notes: All entries are counts or means. Standard deviations appear in parentheses where relevant.

bids gives the total number of observations of WTP elicitations during the BDM phase of the experiment.

revised is the number of subjects who made at least one revision to an earlier round's submission.

wtp is the average of the last submitted willingness to pay (in \$) for each round over all subjects.

safe is the number of times (out of ten rows) that the subject opted for the safer lottery during the Holt-Laury phase.

item tells how many subjects in a given session actually received the item.

price is the value of the randomly drawn price for the BDM mechanism actually resolved for each individual subject.

payment is the amount of monetary compensation that the subjects earned (separate and in addition to any acquisition of the item).

pre_wtp consists of the numerical responses to the question: "How much (in dollars) did you value the item before logging in to start the experiment?" 3 subjects in session 2 did not give numeric responses to this question, so they are omitted.

post_wtp consists of the numerical responses to the question: "How much (in dollars) did you value the item at the end of the experiment, immediately before rolling the dice?" 1 subject in session 1, 5 subjects in session 2, and 1 subject in session 3 did not give numeric responses to this question, so they are omitted.

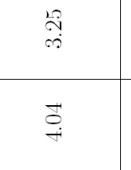
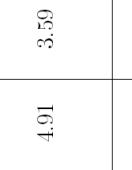


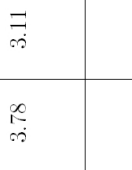
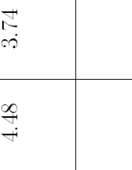
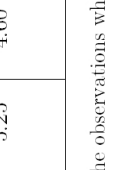
Chapter 2. Testing Distributional Dependence in the Becker-DeGroot-Marschak Mechanism

Table 2.3. Summary of Categorical Survey Responses by Type of Instructions

	response	all	basic	detailed
Q2. Do you like cookies?	I strongly like cookies	16	7	9
	I like cookies	35	16	19
	I somewhat like cookies	13	11	2
	I somewhat dislike cookies	2	0	2
	I dislike cookies	1	0	1
	I strongly dislike cookies	1	0	1
Q4. Did your value for the item change over the course of the experiment?	Increased	10	5	5
	Stayed about the same	32	14	18
	Decreased	19	11	8
	Both increased and decreased	7	4	3
Q7. How did you decide what values to submit each round?	Predetermined or fixed valuation	14	6	8
	Zero or low valuation	8	4	4
	Bimodal, depending on dist.	8	2	6
	Mean, median, or other quartile	9	7	2
	Spike in probability	2	2	0
	Probability (vaguely stated)	5	2	3
	Preference (vaguely stated)	2	2	0
	Hybrid of prob. & preference	10	7	3
	Chasing satisfaction of winning	2	1	1
	Random	4	1	3
Other	4	2	2	
Q9. How do you feel about the outcome of the experiment?	Pleased	28	14	14
	Disappointed	6	4	2
	Excited	11	6	5
	Bored	9	6	3
	Satisfied	38	18	20
	Regretful	4	2	2
	Hungry	15	5	10
Q10. Were the instructions for Phase I of the experiment clear?	Very clear	11	5	6
	Clear	28	12	16
	Somewhat clear	17	11	6
	Somewhat unclear	7	3	4
	Unclear	2	2	0
	Very unclear	3	1	2

Notes: One subject did not respond to the survey, thus $N = 68$. Question 7 called for open, free-form responses, which I categorized *ex post*. Full responses to Question 7 and their categorizations appear in Appendix 2.C. Subjects were allowed to check multiple options for Question 9.

Table 2.4. Descriptive Statistics of BDM Elicitations by Distribution [N = 69]

#	Std. Dev.			Thumbnail	Mean	Min	Max
	#	Thumbnail	Mean				
1	8		4.04	3.25	0	13	
2	9		4.91	3.59	0	14.5	
3	10		4.99	4.60	0	15	
4	11		2.42	2.60	0	9.5	
5	12		3.78	3.11	0	10	
6	13		4.48	3.74	0	14.5	
7	14		5.25	4.60	0	14.5	

Note: Distributions 1, 2, 4, 5, 9, and 14 each appeared twice during the 20 rounds. The statistics reported here reflect only the observations when subjects submitted a valuation for a specified distribution for the *last* time.

Table 2.5. Subjects with Significant Coefficients Regressing WTP on Median

rounds	distributions	subjects with coefficients of <i>Median</i> significantly different from zero			subjects with coefficients of <i>Median</i> significantly less than zero		
		all subjects	basic instructions	detailed instructions	all subjects	basic instructions	detailed instructions
all	all	36 (33)	21 (19)	15 (14)	5 (4)	0 (0)	5 (4)
	low support	19 (15)	13 (12)	6 (3)	1 (1)	1 (1)	0 (0)
	high support	18 (13)	13 (10)	5 (3)	4 (2)	2 (1)	2 (1)
first 10	all	35 (28)	20 (18)	15 (10)	5 (2)	0 (0)	5 (2)
last 10		36 (27)	20 (16)	16 (11)	7 (6)	0 (0)	7 (6)

Notes: Each row summarizes the results from running the regression $WTP = \beta_0 + \beta_1 \cdot Median + \varepsilon$ separately for each of the 69 subjects, restricting the sample to the appropriate rounds or support depending on the row (for a total of 345 regressions). The counts given in the table are the number of subjects for whom the coefficient β_1 is significantly different from zero at a 5% (1%) level. The results are virtually identical when *Mean* is used as the regressor instead of *Median*.





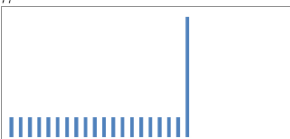
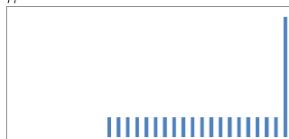








Chapter 2. Testing Distributional Dependence in the Becker-DeGroot-Marschak Mechanism

Table 2.6. Pairwise Distributional Tests: Change in Location of Mass [N = 69]

Distributions Compared		p-value for test of means		
		all subjects	basic instructions	detailed instructions
#3 	#4 	0.0141**	0.0077***	0.2422
#5 	#6 	0.0322**	0.0226**	0.2484
#7 	#8 	0.0062***	0.0015***	0.2223
#9 	#10 	0.4509	0.2054	0.7302
#11 	#12 	0.0030***	0.0008***	0.2088
#13 	#14 	0.1395	0.0315**	0.6302

Notes: Reported p-values give the probability from the one-tailed test that the mean of the elicitation from the distribution with more mass to the left is greater than the mean of the elicitation from the matched paired distribution with more mass to the right. Distributions 4, 5, 9, and 14 each appeared twice during the 20 rounds. The tests reported here reflect only the observations when subjects submitted a valuation for a specified distribution for the *last* time. The symbols *, **, and *** indicate significance at the 10%, 5%, and 1% levels respectively.

Table 2.7. Pairwise Distributional Tests: Change in Support [N = 69]

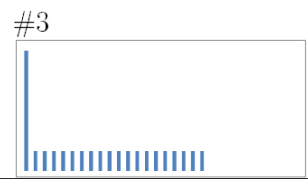
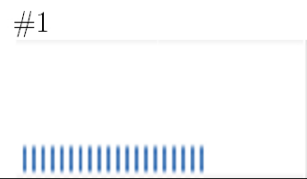
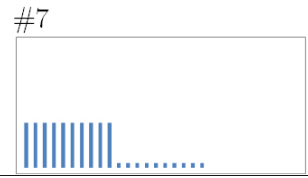

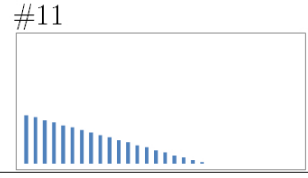

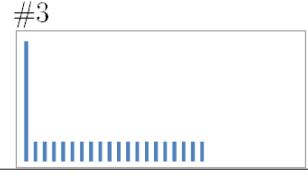
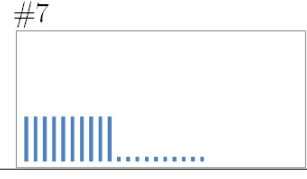
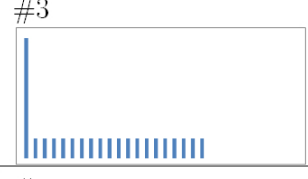
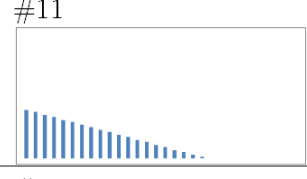
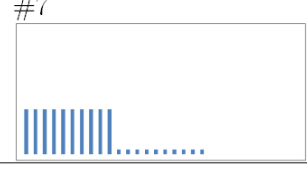
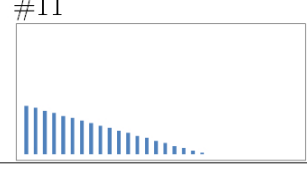
Distributions Compared		p-value for test of means		
		all subjects	basic instructions	detailed instructions
#1 	#2 	0.0043***	0.0036***	0.1646
#3 	#5 	0.0009***	0.0010***	0.0853*
#4 	#6 	0.0048***	0.0075***	0.0991*
#7 	#9 	0.0000***	0.0000***	0.0539*
#8 	#10 	0.0804*	0.0352**	0.4252
#11 	#13 	0.0001***	0.0005***	0.0264**
#12 	#14 	0.0147**	0.0120**	0.1867

Notes: Identical to the notes of Table 6, with the amendment that distributions 1 and 2 also appeared twice for each subject over the course of the 20 rounds of the BDM phase (in addition to distributions 4, 5, 9, and 14)

Chapter 2. Testing Distributional Dependence in the Becker-DeGroot-Marschak Mechanism

Table 2.8. Pairwise Distributional Tests: Change in Distribution Shape [N = 69]





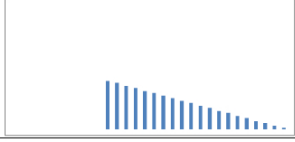

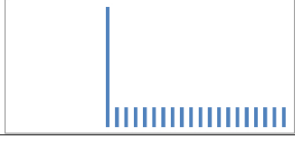

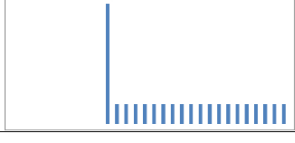
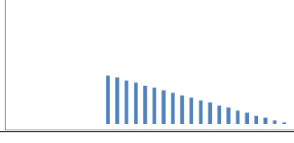

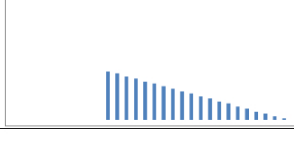
Panel A. Mass to the Left, Low Support

Distributions Compared		p-value for test of means		
		all subjects	basic instructions	detailed instructions
		0.0847*	0.0993*	0.2566
		0.1499	0.0955*	0.4040
		0.0356**	0.0628*	0.1441
		0.3215	0.4241	0.3286
		0.6469	0.5440	0.6599
		0.8169	0.6344	0.8097

Notes: Reported p-values give the probability from the one-tailed test that the mean of the elicitations from the distribution with more mass to the left is greater than the mean of the elicitations from the matched paired distribution with more mass to the right. Distributions 1, 2, 4, 5, 9, and 14 each appeared twice during the 20 rounds. The tests reported here reflect only the observations when subjects submitted a valuation for a specified distribution for the *last* time. The symbols *, **, and *** indicate significance at the 10%, 5%, and 1% levels respectively.

Chapter 2. Testing Distributional Dependence in the Becker-DeGroot-Marschak Mechanism


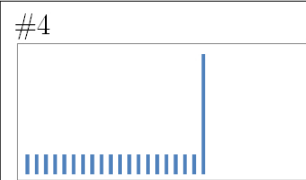

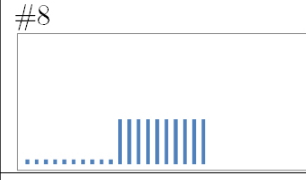

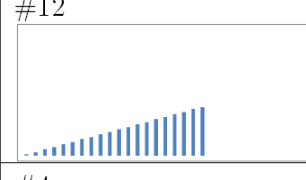

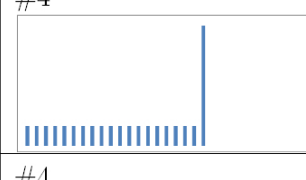
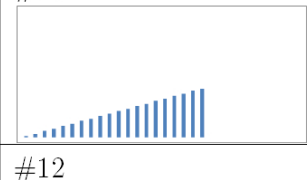
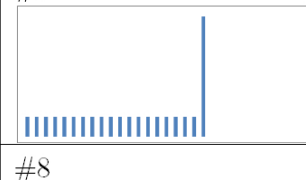
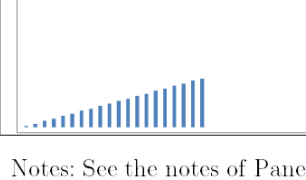
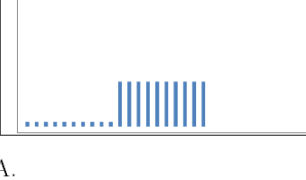
Panel B. Mass to the Left, High Support

Distributions Compared		p-value for test of means		
		all subjects	basic instructions	detailed instructions
#5 	#2 	0.1827	0.1299	0.4357
#9 	#2 	0.4938	0.3241	0.6638
#13 	#2 	0.2568	0.1543	0.5374
#5 	#9 	0.1692	0.1979	0.2867
#5 	#13 	0.4028	0.4632	0.4038
#9 	#13 	0.7530	0.7635	0.6219

Notes: See the notes of Panel A.

Chapter 2. Testing Distributional Dependence in the Becker-DeGroot-Marschak Mechanism


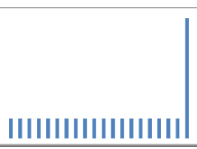



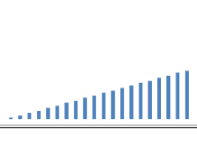
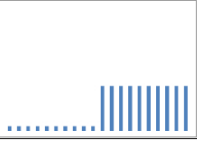

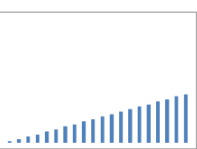

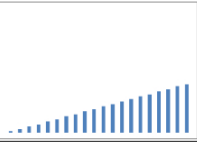
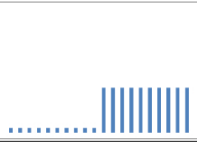
Panel C. Mass to the Right, Low Support

Distributions Compared		p-value for test of means		
		all subjects	basic instructions	detailed instructions
#1 	#4 	0.1772	0.1016	0.4645
#1 	#8 	0.0800*	0.0620*	0.3056
#1 	#12 	0.1701	0.0474**	0.5992
#8 	#4 	0.6620	0.5762	0.6526
#12 	#4 	0.4897	0.6215	0.3729
#12 	#8 	0.3200	0.5486	0.2262

Notes: See the notes of Panel A.

Chapter 2. Testing Distributional Dependence in the Becker-DeGroot-Marschak Mechanism

Panel D. Mass to the Right, High Support

Distributions Compared		p-value for test of means		
		all subjects	basic instructions	detailed instructions
#2 	#6 	0.1673	0.2010	0.2915
#2 	#10 	0.4590	0.3676	0.5917
#2 	#14 	0.3250	0.2055	0.6014
#10 	#6 	0.2063	0.3140	0.2345
#14 	#6 	0.3112	0.4949	0.2234
#14 	#10 	0.6302	0.6809	0.4940

Notes: See the notes of Panel A.

Table 2.9. p-values for Tests of Means Between Elicitations from Basic Instruction and Elicitations from Detailed Instructions, by Round and Timing of Elicitations

round	first	last
1	0.0327**	0.0071***
2	0.0406**	0.0091***
3	0.4858	0.0998*
4	0.0756*	0.3180
5	0.0389**	0.0180**
6	0.0254**	0.0209**
7	0.0995*	0.0175**
8	0.1352	0.1538
9	0.2801	0.1755
10	0.9676	0.4546
11	0.5295	0.2834
12	0.7827	0.8337
13	0.1115	0.1618
14	0.0963*	0.0282**
15	0.1193	0.0631*
16	0.0746*	0.0783*
17	0.9915	0.6717
18	0.2531	0.1310
19	0.1053	0.1151
20	0.7956	0.4165

Notes: For each round, the reported p-values give the probability from the two-tailed test that the mean of the elicitation from subjects who viewed detailed instructions has the same value as the mean of the elicitation from subjects who viewed basic instructions. The column “last” gives these test values for the relevant means over the *last* elicitation in a given round, and the p-values reported thus give an idea of the statistical significance of differences of means displayed visually in Figure 3. The column labeled “first” gives these test values for the relevant means over the *first* elicitation in a given round. The symbols *, **, and *** indicate significance at the 10%, 5%, and 1% levels respectively.

Appendix 2.A Experimental Instructions and Materials

Overview of the Experiment

This is an experiment in the economics of decision-making. Research foundations have provided funds for conducting this research. The experiment should take less than an hour. At the end of the experiment you will be paid privately by the XLab staff in the form of a check.

At this time, **you have been given \$15** in an electronic account. This amount may change over the course of the experiment as a result of both the decisions you make and also chance.

In addition, you will notice on your desk a gift certificate from CREAM (Cookies Rule Everything Around Me), which is a popular new establishment at the intersection of Telegraph and Channing that sells ice cream and cookies. **The gift certificate is redeemable for one dozen freshly baked cookies.** Depending on your decisions and also on chance, you may be able to obtain this gift certificate, which is referred to as "the item" in the instructions that follow.

Details regarding the decisions you will be asked to make are provided on the following pages.

Rules

Please remain silent during the experiment. Your participation in the experiment and any information about your earnings will be kept strictly confidential. Your payments receipt and participant form are the only places in which your name and social security number are recorded, and these will never be viewed or recorded by the experimenter.

Login

If there are no further questions, I will collect the consent forms and hand out both the five-digit login codes and hard copies of the instructions. You may then login and begin the experiment.

login code:

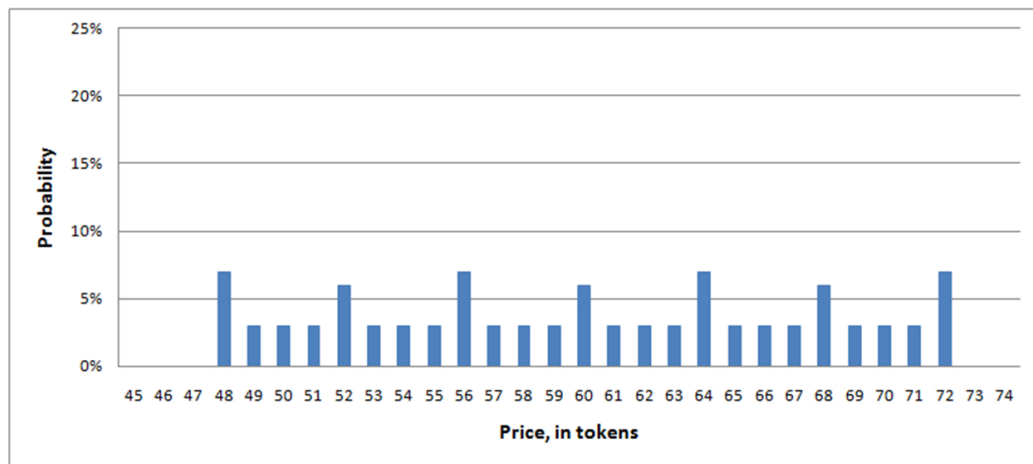
Figure 2.A.1. Introductory Screen

Chapter 2. Testing Distributional Dependence in the Becker-DeGroot-Marschak Mechanism

Phase I Instructions [First page common to both instruction types.]

This portion of the experiment is divided into 20 independent rounds. At the conclusion of the experiment, **one** of the 20 rounds will be selected at random, and only that one will count for determining whether you buy the item (depending on the decision you made in that round) and, if so, how much you pay for it.

In each of these 20 rounds, you will be faced with a distribution of possible prices, depicted in graphical form. Some of the distributions place greater weight on particular possible prices, making them more likely to occur. The heights of the bars in the graph tell how likely each possible price is for that distribution. For example, consider the following distribution of random prices:



There is a 7% chance the price will be 48 tokens, a 3% chance the price will be 49 tokens, a 3% chance the price will be 50 tokens, a 3% chance the price will be 51 tokens, a 6% chance the price will be 52 tokens, and so on. There is a 0% chance that the price will be 45, 46, 47, 73, or 74 tokens—those prices are not possible for this distribution.

The experiment has more rounds than distributions, so you will see some of the distributions of possible prices more than once.

In each round—after a brief time for consideration of the relevant distribution of possible prices—you will submit your willingness to pay (*WTP*) for the item at your desk.

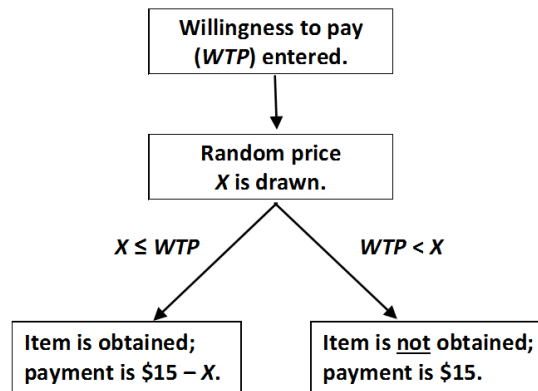
For the round that is chosen to count, a random price X will be drawn from the distribution of possible prices that corresponds to that round.

- If your stated willingness to pay for the item in the task in question is greater than or equal to the randomly drawn price (i.e., if $WTP \geq X$), then you purchase the item **for the randomly drawn price X** —not your submitted willingness to pay (unless X happens to equal your submitted WTP exactly). You will receive the item at the conclusion of the experiment, and the amount X will be deducted from your endowment.
- If your stated willingness to pay for the item in the task in question is strictly less than the randomly drawn price (i.e., if $WTP < X$), then you cannot purchase the item for price X or any other price. You will not receive the item, and nothing will be deducted from your endowment.

Chapter 2. Testing Distributional Dependence in the Becker-DeGroot-Marschak Mechanism

[Second page of the instructions. The unboxed text appeared in both sets of instructions, while the boxed text and figure were included only in the detailed instructions. The boxes themselves did not appear in the actual detailed instructions.]

That is, for the round that is chosen to count, the payoffs will be determined at the end of the experiment as follows:



For example, consider the distribution above, and imagine that you had submitted a willingness to pay of 61 tokens. If the randomly drawn price were, say, 53 tokens, you would receive the item and pay 53 tokens—not 61 tokens. On the other hand, if the randomly drawn price were, say, 64 tokens, you would not receive the item and you would not pay anything.

Notice that given this set of rules, **there is every incentive for you to report your willingness to pay truthfully**. Consider again the example above: if your true value for the item were 61 tokens there would be no benefit from submitting any other value—either higher or lower—and in fact there may be a disadvantage from doing so. If you submitted a willingness to pay of 57 tokens and the randomly drawn price were 59 tokens, you would not receive the item even though you would have actually been willing to pay 59 tokens. Contrariwise, if you submitted a willingness to pay of 67 tokens and the randomly drawn price were 63 tokens, you would receive the item and pay 63 tokens—more than your true value of 61 tokens. Submitting your true value as your willingness to pay avoids both of these potential pitfalls.

Keep in mind that in any of the coming rounds you are always free to enter a willingness to pay of **any** amount—whether or not it is above or below all of the possible random prices. Your submitted *WTP* will still be compared to the randomly drawn price *X* as described above, and you can never possibly pay anything other than the randomly drawn price. **Regardless of whether your true willingness to pay falls above or below all of the possible random prices there would still be no reason to submit any other value than your true willingness to pay.**

Since each round has an equal likelihood of being selected, you should approach each round with equal care and consideration.

At any time before the end of Phase I you are free to return to earlier rounds that you have already completed and revise your willingness to pay if you so choose. Just click on the number of the earlier round (under the graph) and submit a new willingness to pay.

Chapter 2. Testing Distributional Dependence in the Becker-DeGroot-Marschak Mechanism

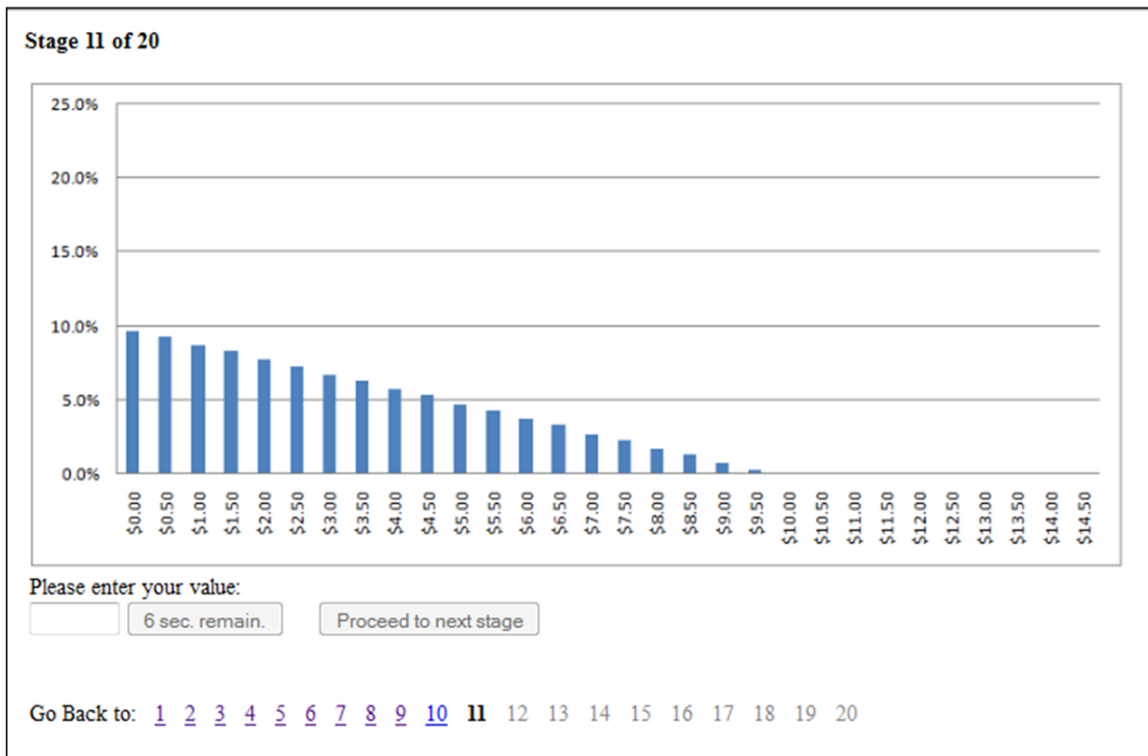


Figure 2.A.2. Screenshot of a Round of the BDM Phase

Chapter 2. Testing Distributional Dependence in the Becker-DeGroot-Marschak Mechanism

Phase II

For each row in the table that follows, indicate which of the lotteries you would prefer to have by clicking the appropriate button.

At the conclusion of the experiment, **one** of these ten rows will be selected at random and your chosen lottery in that row will be played. The amount earned will be added your payment from the first phase of the experiment.


1	<input type="radio"/> 10% \$2.00 and 90% \$1.60	<input type="radio"/> 10% \$3.85 and 90% \$0.10
2	<input type="radio"/> 20% \$2.00 and 80% \$1.60	<input type="radio"/> 20% \$3.85 and 80% \$0.10
3	<input type="radio"/> 30% \$2.00 and 70% \$1.60	<input type="radio"/> 30% \$3.85 and 70% \$0.10
4	<input type="radio"/> 40% \$2.00 and 60% \$1.60	<input type="radio"/> 40% \$3.85 and 60% \$0.10
5	<input type="radio"/> 50% \$2.00 and 50% \$1.60	<input type="radio"/> 50% \$3.85 and 50% \$0.10
6	<input type="radio"/> 60% \$2.00 and 40% \$1.60	<input type="radio"/> 60% \$3.85 and 40% \$0.10
7	<input type="radio"/> 70% \$2.00 and 30% \$1.60	<input type="radio"/> 70% \$3.85 and 30% \$0.10
8	<input type="radio"/> 80% \$2.00 and 20% \$1.60	<input type="radio"/> 80% \$3.85 and 20% \$0.10
9	<input type="radio"/> 90% \$2.00 and 10% \$1.60	<input type="radio"/> 90% \$3.85 and 10% \$0.10
10	<input type="radio"/> 100% \$2.00 and 0% \$1.60	<input type="radio"/> 100% \$3.85 and 0% \$0.10


Figure 2.A.3. Screenshot of the Holt-Laury Phase


Chapter 2. Testing Distributional Dependence in the Becker-DeGroot-Marschak Mechanism


Payoff Determination

So that you can be absolutely confident that the results of the experiment are fair and that the outcomes aren't predetermined, you will roll a set of seven dice that—together with your decisions in Phase I and Phase II—will decide whether or not you receive the item and how much money you will earn.

 The black 20-sided die will determine which of the 20 rounds of Phase I will count.

 The red, blue, and green 10-sided dice will, added together, yield a random decimal number between 0.001 and 1.000 (the sum 0.000 represents the number 1.000). This random decimal number will determine the randomly drawn price X depending on the distribution given by the randomly selected round.

 The black 10-sided die will determine which of the 10 rows of Phase II will count ("0" represents the tenth row).

 The white and gray 10-sided dice will, added together, yield a random percentile (the sum 00 represents 100%). If the value of the random percentile is equal to or less than the percentage listed for the higher outcome in the lottery you have chosen in the randomly selected row, you will receive the higher amount. Otherwise, you will receive the lower amount.

At this time, please do the following:

1. Write the number of your terminal (x01, x02, or similar) on the slip of paper with the five-digit login code.
2. Take the slip of paper and the item from your desk and give them to the experimenter at the back of the room. Please wait silently in line until the experimenter is able to attend you.
3. Roll the dice and learn the outcome of the experiment. Depending on your decisions and the dice rolls, the item may or may not be returned to you.

...

Now that you have rolled the dice and your payoff has been determined, please take a few minutes to complete the following survey silently while the XLab staff prepares your payment.

[Link to Survey](#) **[Do not begin the survey until after you have rolled the dice.]**

Figure 2.A.4. Screenshot of the Payoff-Determination Briefing

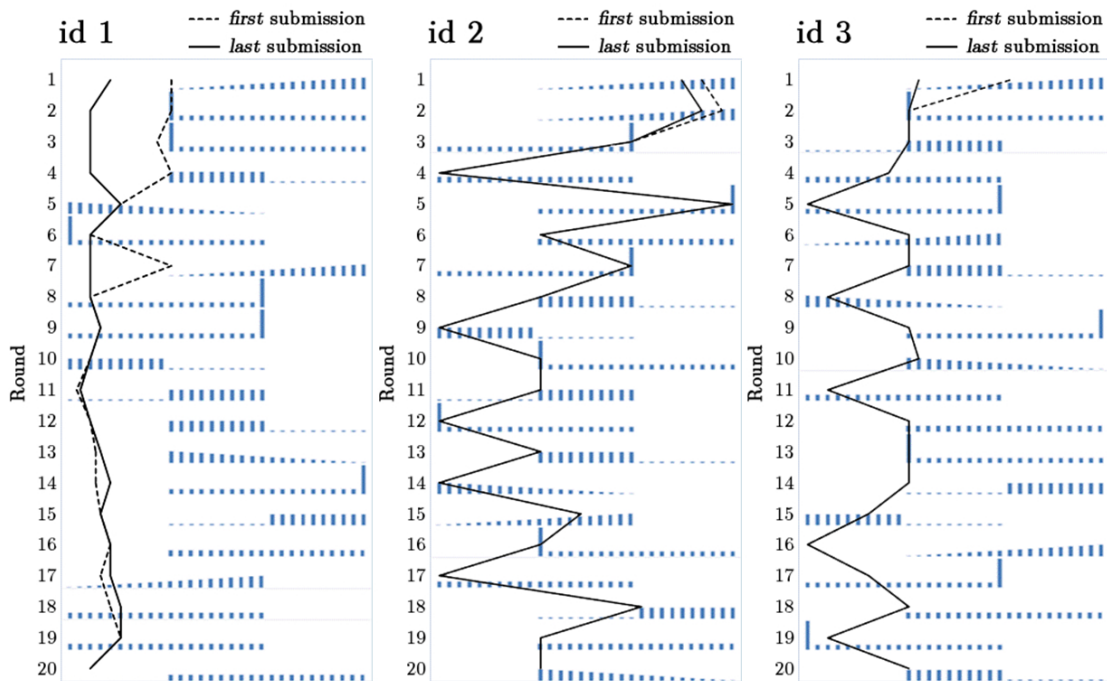
Chapter 2. Testing Distributional Dependence in the Becker-DeGroot-Marschak Mechanism

Survey Questions

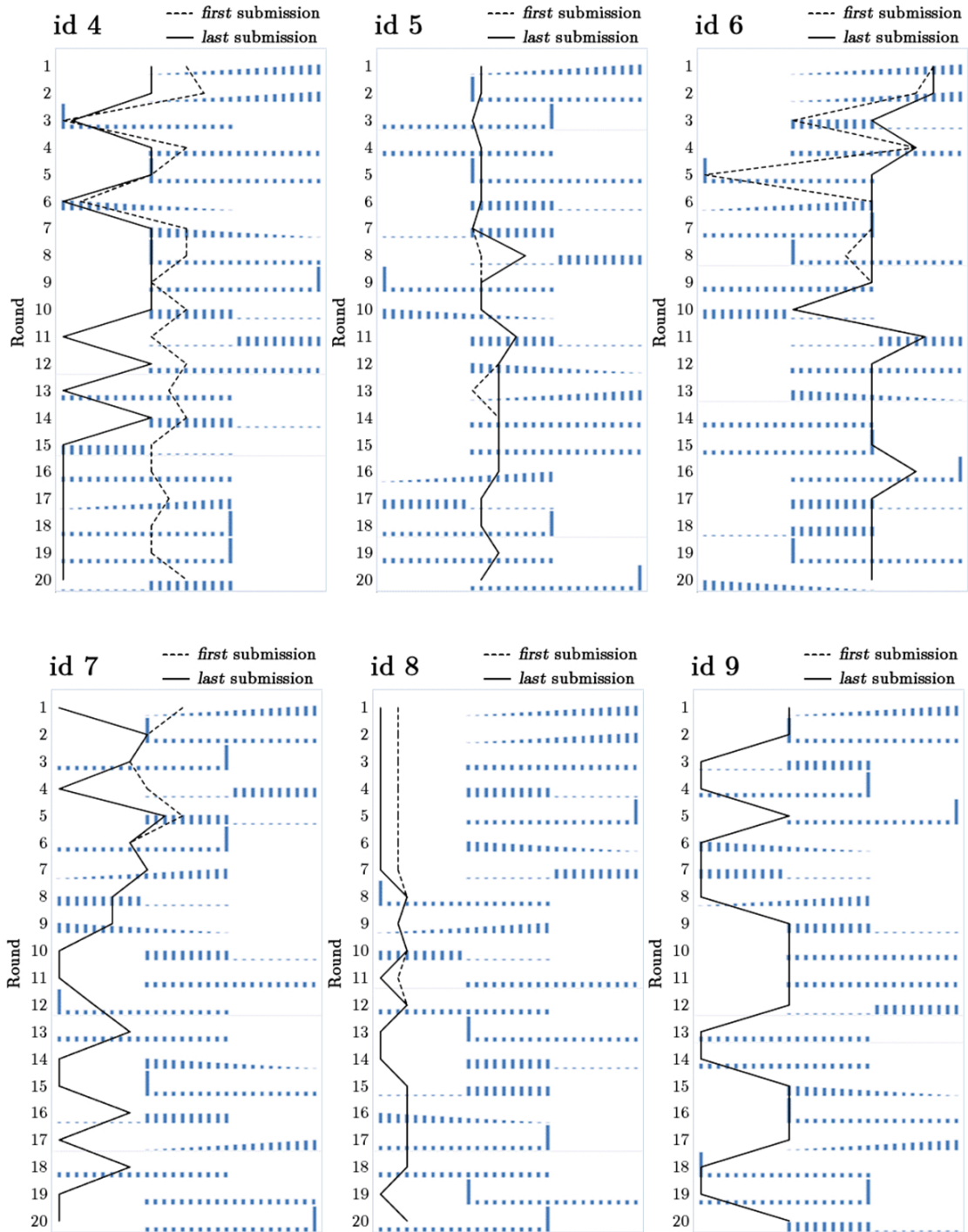
1. What is your five-digit login code?
2. Do you like cookies? (select one)
 - I strongly dislike cookies
 - I dislike cookies
 - I somewhat dislike cookies
 - I somewhat like cookies
 - I like cookies
 - I strongly like cookies
3. How much (in dollars) did you value the item before logging in to start the experiment?
4. Did your value for the item change over the course of the experiment? (select one)
 - My value for the item decreased over the course of the experiment
 - My value for the item stayed about the same over the course of the experiment
 - My value for the item increased over the course of the experiment
 - My value for the item both increases and decreased sometime during the experiment
5. Why did your value change over the course of the experiment, or why didn't it change?
6. How much (in dollars) did you value the item at the end of the experiment, immediately before rolling the dice?
7. How did you decide what values to submit each round?
8. If you went back and revised earlier submissions, what prompted you to do so? If not, why not?
9. How do you feel about the outcome of the experiment? (select as many as apply)
 - Hungry
 - Disappointed
 - Regretful
 - Bored
 - Satisfied
 - Pleased
 - Other (please specify)
10. Were the instructions for Phase I clear? (select one)
 - Very unclear
 - Unclear
 - Somewhat unclear
 - Somewhat clear
 - Clear
 - Very clear

Appendix 2.B Visualizations of Experimental Data

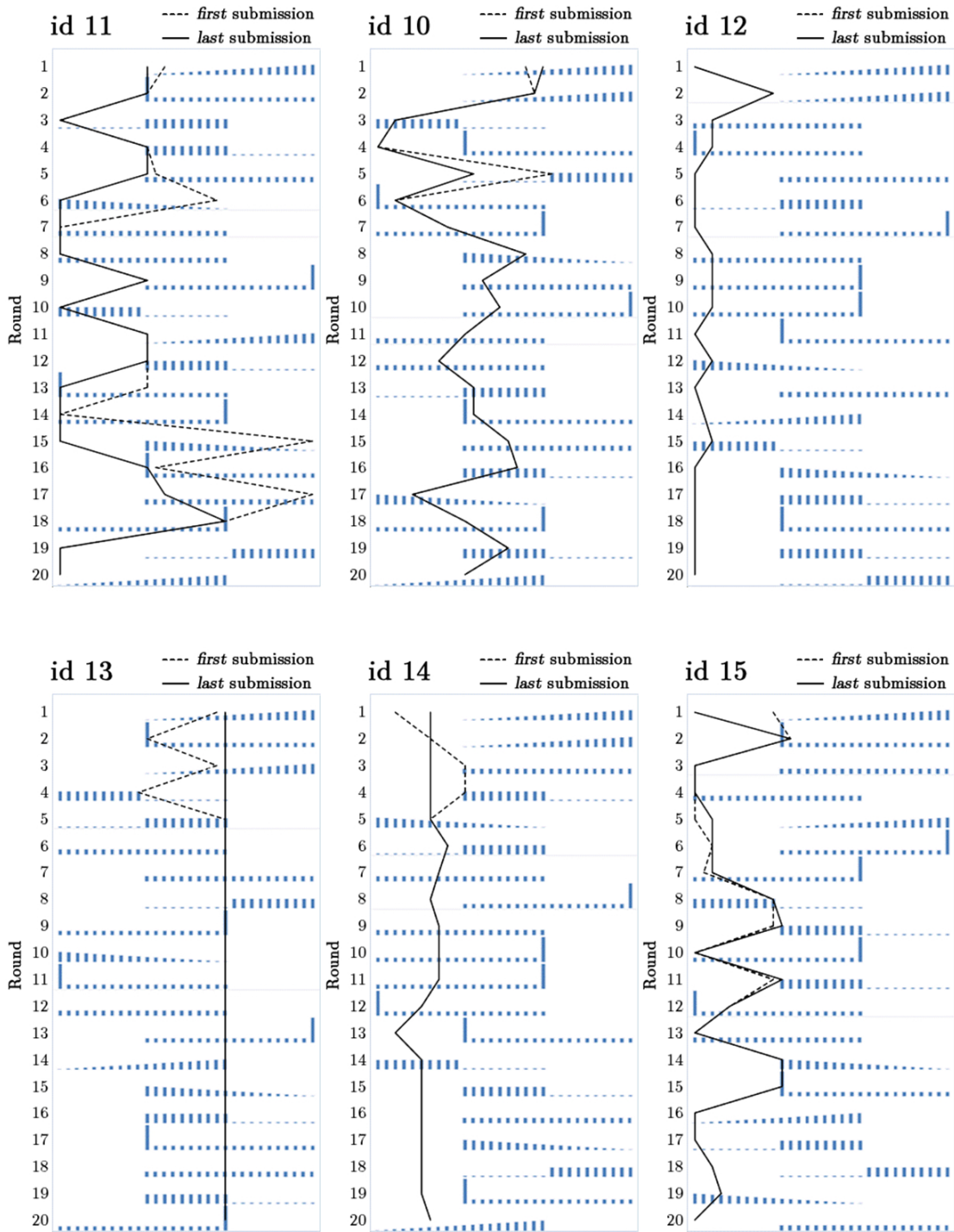
Notes: Where the first and last elicitations for a given round are the same, only the last one is given. The mass points for the distributions occur at intervals of 50¢. The low-support distributions range from \$0.00 (at the left edge of the graph) to \$9.50. The high-support distributions range from \$5.00 to \$14.50 (at the right edge of the graph).



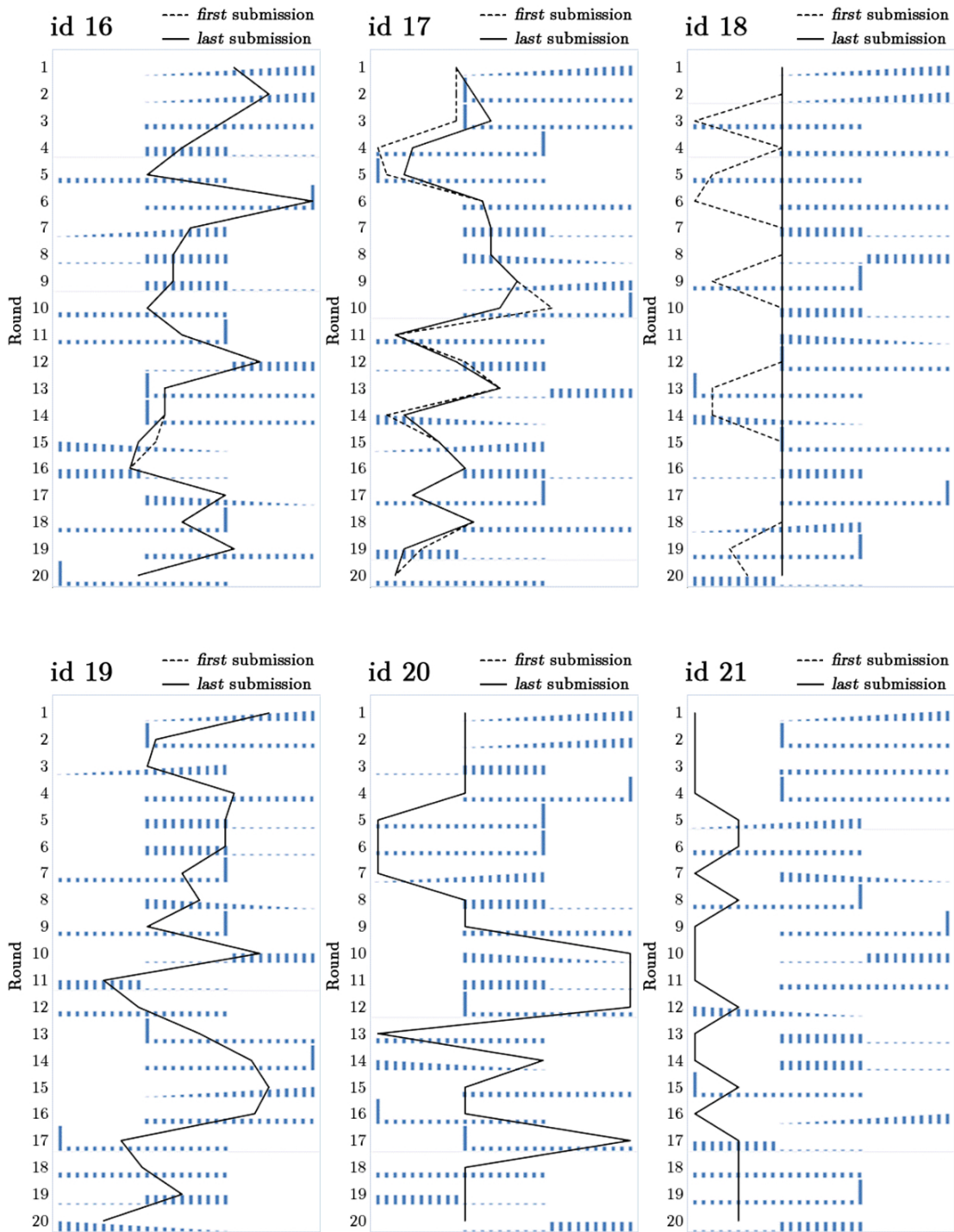
Chapter 2. Testing Distributional Dependence in the Becker-DeGroot-Marschak Mechanism



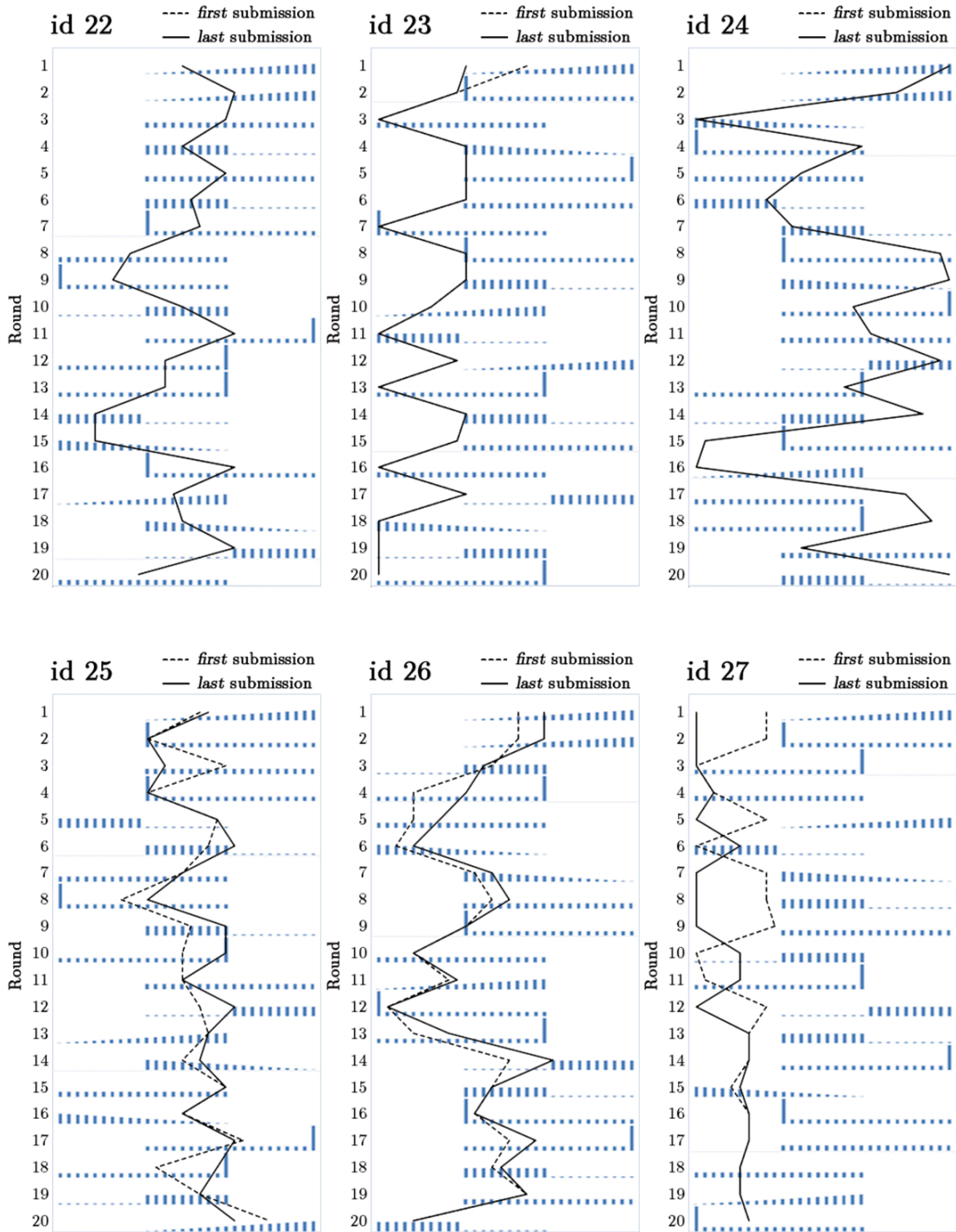
Chapter 2. Testing Distributional Dependence in the Becker-DeGroot-Marschak Mechanism



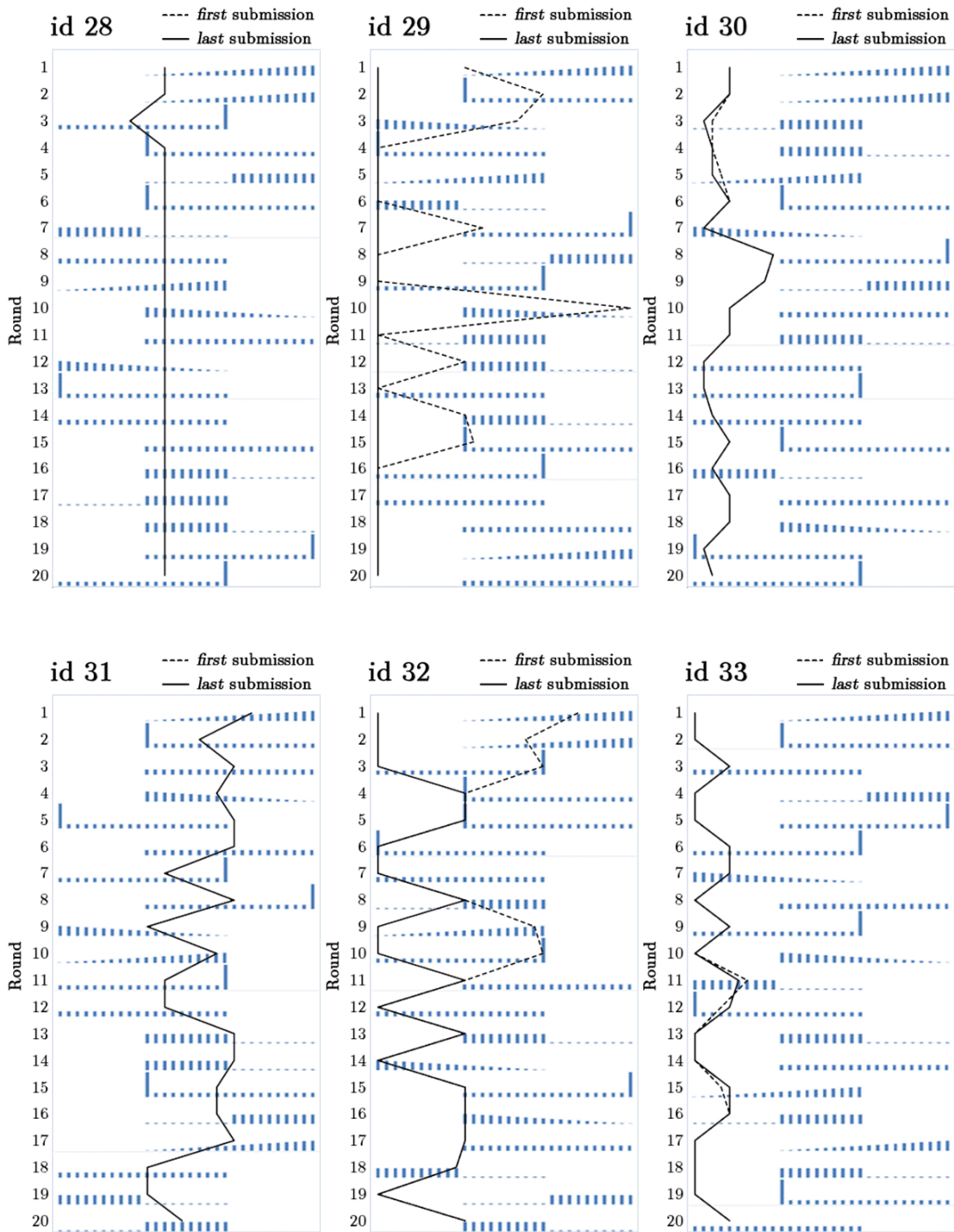
Chapter 2. Testing Distributional Dependence in the Becker-DeGroot-Marschak Mechanism



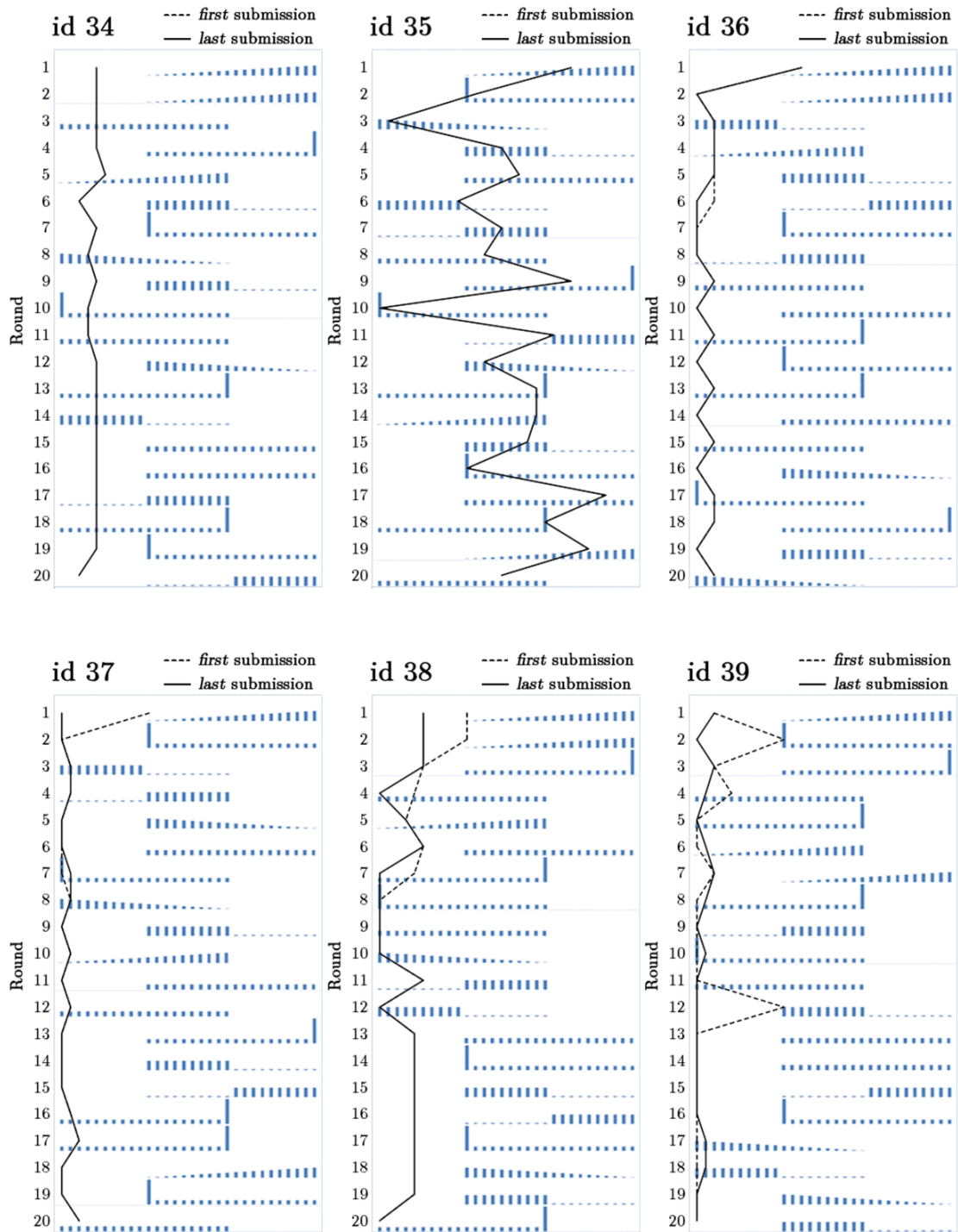
Chapter 2. Testing Distributional Dependence in the Becker-DeGroot-Marschak Mechanism



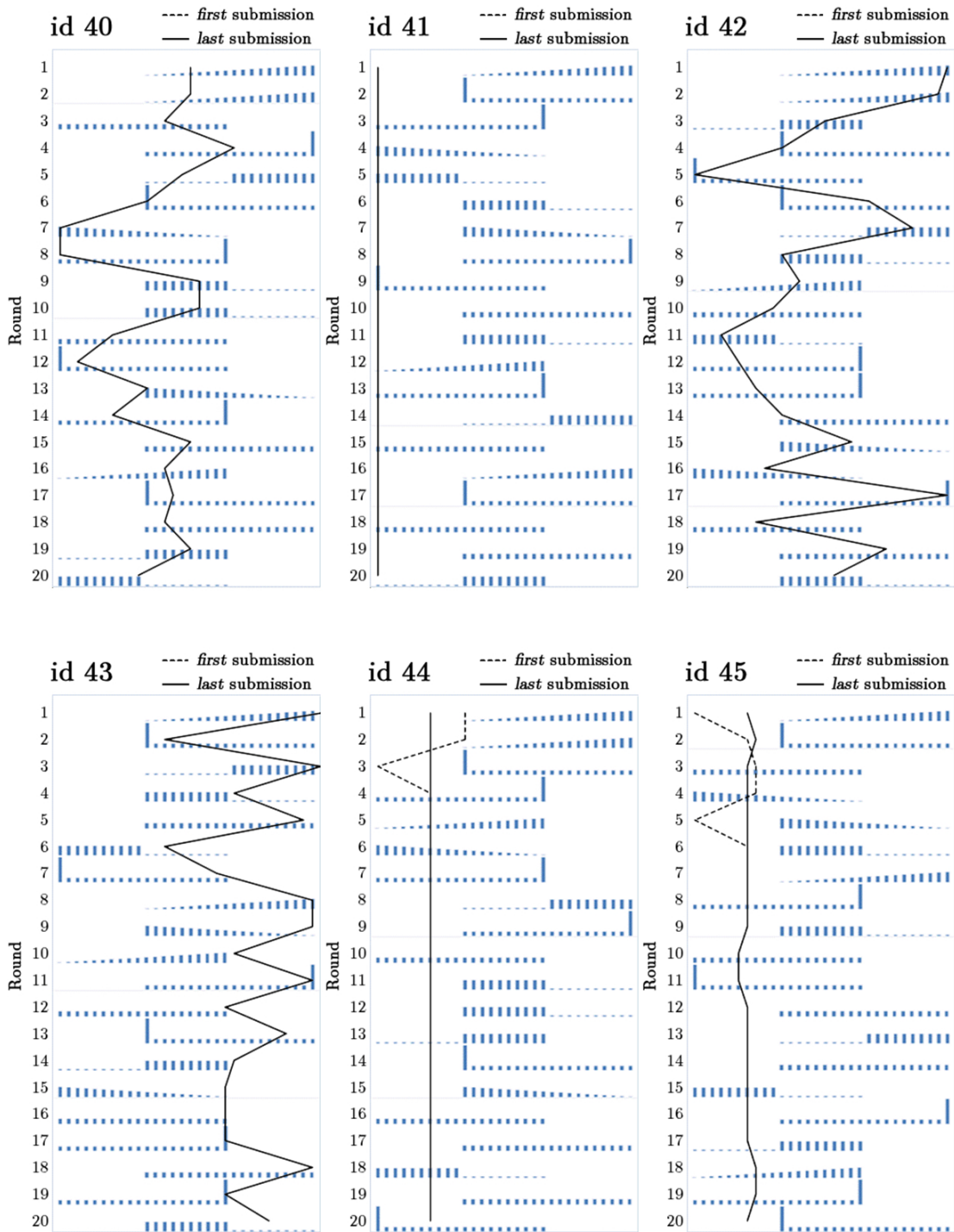
Chapter 2. Testing Distributional Dependence in the Becker-DeGroot-Marschak Mechanism



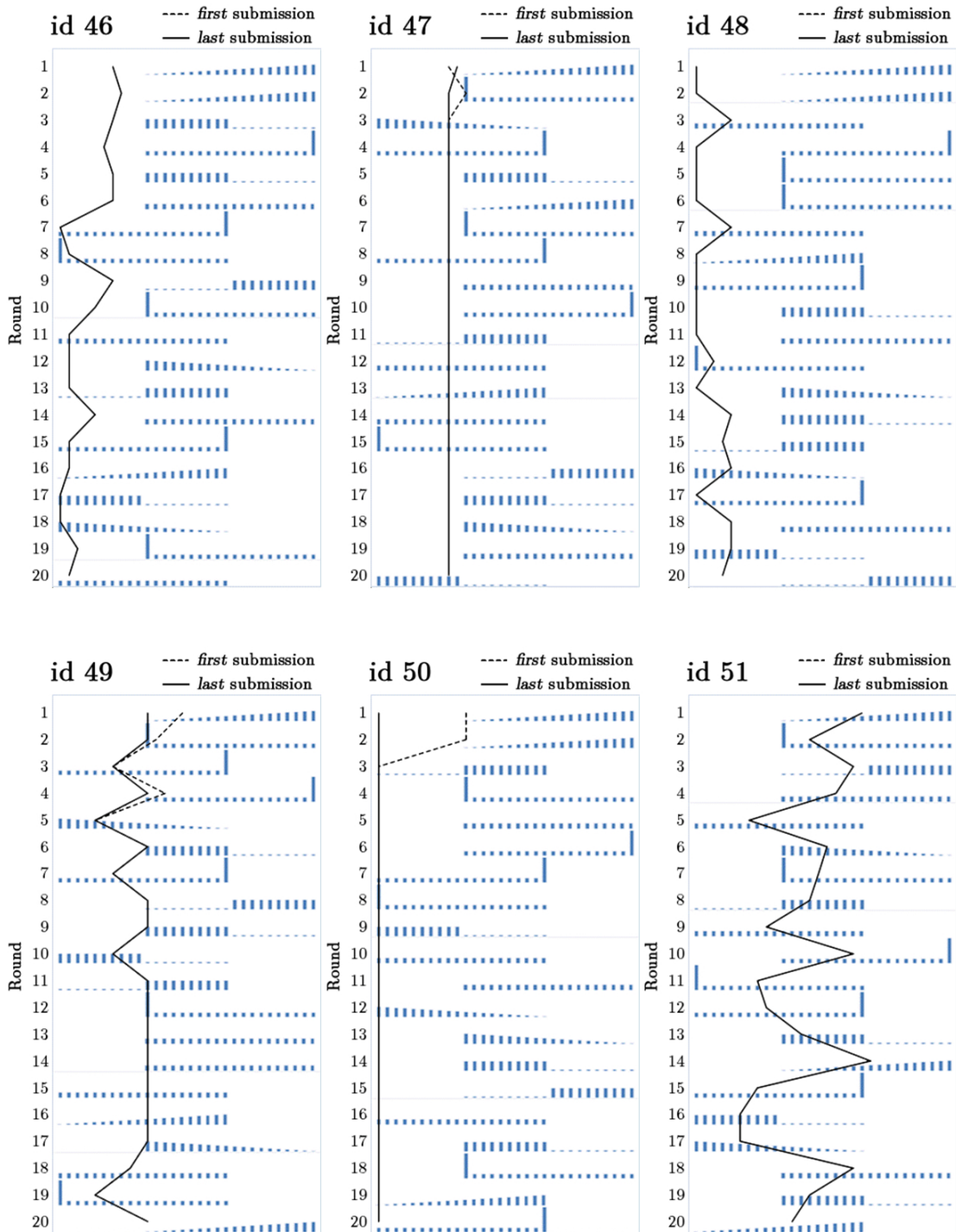
Chapter 2. Testing Distributional Dependence in the Becker-DeGroot-Marschak Mechanism



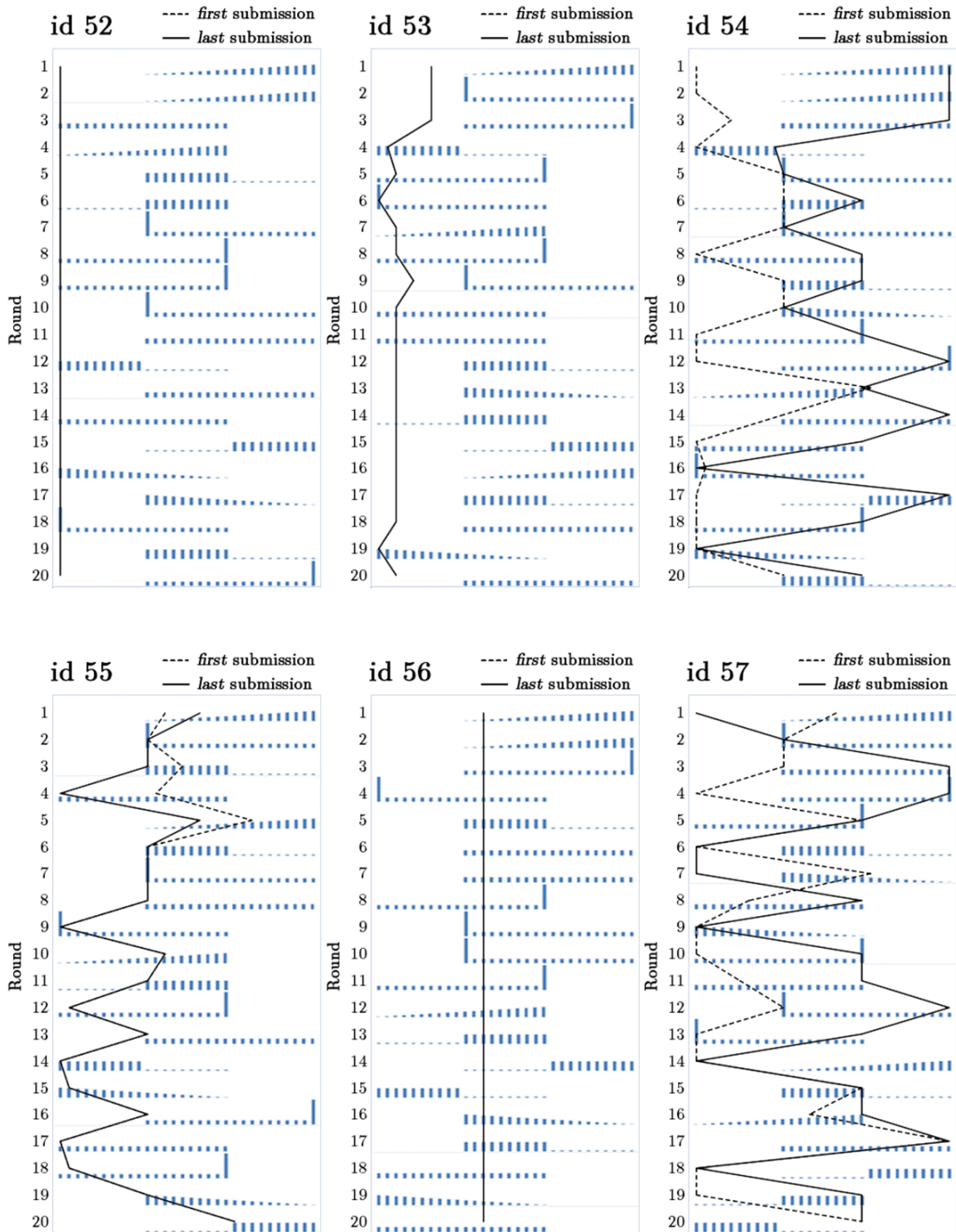
Chapter 2. Testing Distributional Dependence in the Becker-DeGroot-Marschak Mechanism



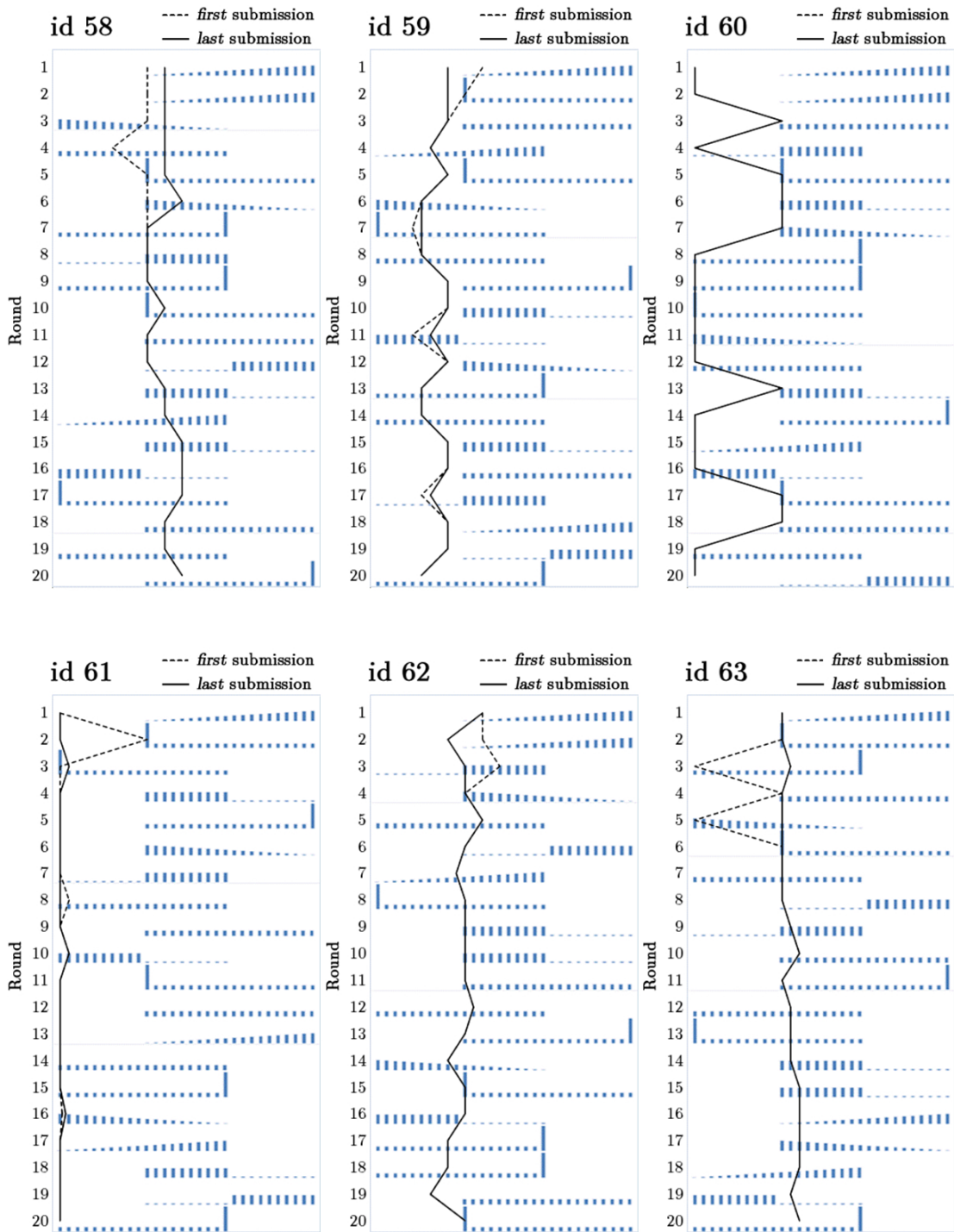
Chapter 2. Testing Distributional Dependence in the Becker-DeGroot-Marschak Mechanism



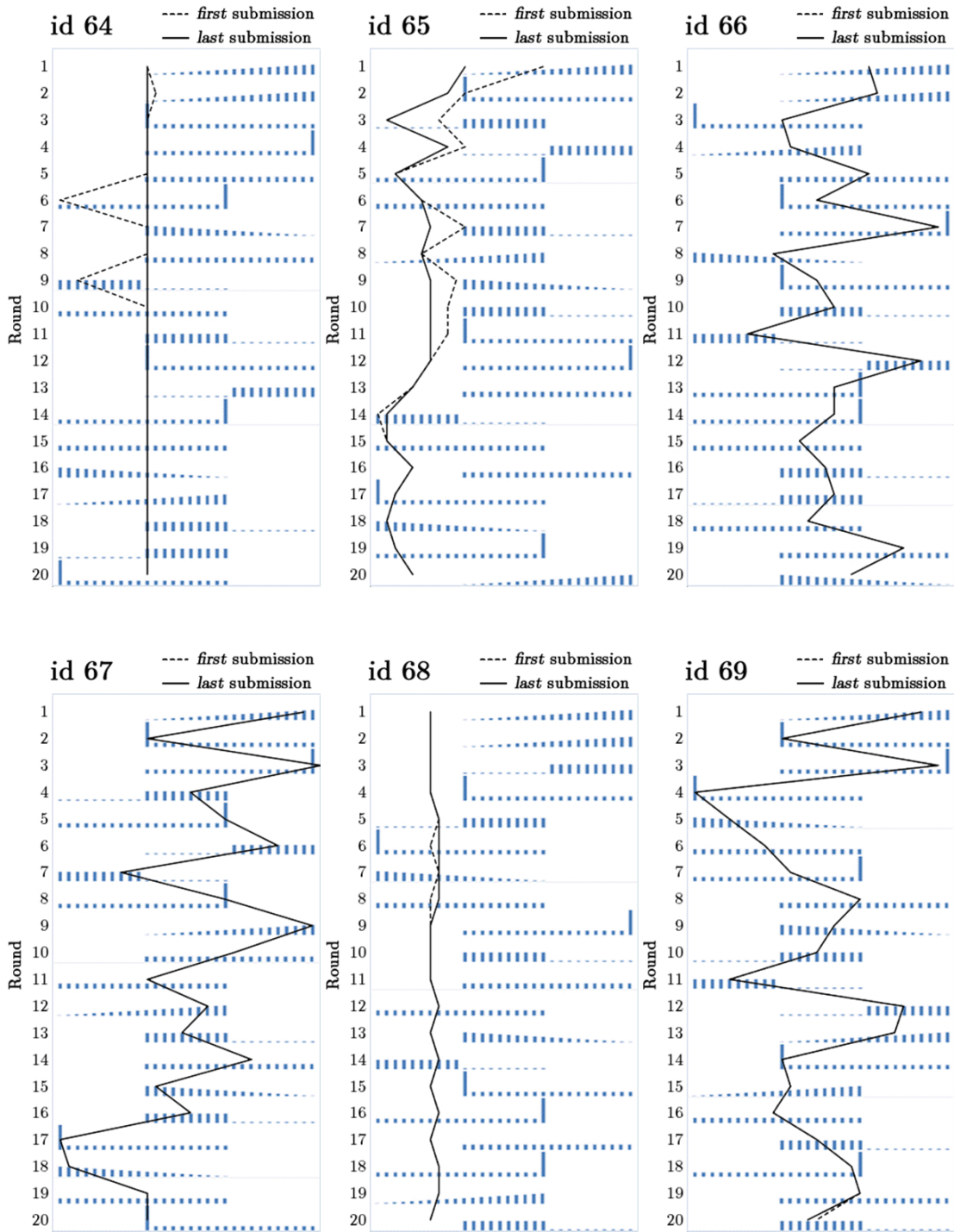
Chapter 2. Testing Distributional Dependence in the Becker-DeGroot-Marschak Mechanism



Chapter 2. Testing Distributional Dependence in the Becker-DeGroot-Marschak Mechanism



Chapter 2. Testing Distributional Dependence in the Becker-DeGroot-Marschak Mechanism



Appendix 2.C Subjects' ex post Explanations of Their Submissions

Note: Det. = 1 indicates detailed instructions, while Det. = 0 indicates basic instructions. The comments are the subjects' responses to the survey question: "How did you decide what values to submit each round?"

Id	Det.	Category	Comment
1	1	random	randomly
2	0	spike	I usually picked the most probable price given the distribution.
3	0	predetermined	I looked at the probabilities, but I didn't not want to pay more than around \$5 for the giftcard so that was my reservation point.
4	1	low	I chose the lowest value in order to have the highest chances of getting a higher value.
5	1	predetermined	I just stuck with the maximum I was willing to pay for a batch of cookies from CREAM.
6	0	hybrid	The distribution on the graph/what I believed the item to be worth.
7	0	hybrid	I looked at the odds and tried to get somewhere around a 50/50 chance of getting the CREAM card. If the 50/50 value was above \$5, I thought I would be better off with just cash. If it was at \$5 or less, I considered it.
8	1	hybrid	Well I didn't want to pay a whole lot for it, I'd rather have actual money, I can make my own cookies, but if it was an ok chance, I'd take a dozen cheap cookies!
9	1	bimodal	I submitted either \$5 or \$0.00 for each round based on probability.
10	0	other	if the possibility price is high, i submitted the cheaper price and vice versa.
11	0	predetermined	Bases on how much i wanted the item.
12	1	bimodal	If there were some chances in one or less than one dollar, I chose one dollar. If not, I submitted the value 0.
13	1	probability	Look at the probability of each amount and think!
14	0	hybrid	I was willing to take risk if the odds were in my favor.
15	0	zero	I chose the smallest number possible because I wanted the \$15 rather than the gift card
16	1	satisfaction	by looking at the probability chart. Even though I only valued the card at 5 dollars, I thought receiving the cream card would be more satisfying than 15 dollars
17	1	mean	I chose the possible price by finding the average of each round.
18	0	predetermined	I kept it at what I thought the card was valued. If I chose a value more than what I thought it was worth, it would not have been a good deal for me.
19	0	median	i tried to maintain the 50% probability to purchase the item.
20	1	other	I picked the one that had the lowest probability in each distribution. When the probability was the same across all values, I tried to choose the same values (0, 5, or 14.50).
21	1	bimodal	It depended on how favorable the distribution was. If there were no values at or below \$2.50 in terms of probability, than I submitted 0. If there were some that were less than \$2.50, than I submitted \$2.50.

Chapter 2. Testing Distributional Dependence in the Becker-DeGroot-Marschak Mechanism

Id	Det.	Category	Comment
22	0	median	The median.
23	0	hybrid	How much I would like to pay for the cookies based on the chance that they would be that price.
24	1	random	random
25	1	random	I randomly chose.
26	0	1 st quartile	I aimed for around 25% chance each round. It's not an item I would especially want, but 25% chance seemed acceptable.
27	0	hybrid	Lowest possible values.. for those which have higher chances at the lower ends, i set myself a \$3 ceiling.
28	1	predetermined	I thought that it was reasonable to pay \$0.50 for each cookie.
29	1	zero	I am lactose intolerant. The products at CREAM would make me very sick. I have no use for the item, so I submitted the value of \$0 every round.
30	0	hybrid	Looking at the probabilities and determining whether it was worth the money.
31	0	median	depending on the distribution, roughly 50% of winning the item, or around \$10
32	1	hybrid	Values with the minimum amount with the highest chance
33	1	bimodal	I pretty much put \$0.00 whenever the distribution was at \$5 or more, since I didn't want to pay more than \$5 but also didn't feel like coming up with an amount I would pay, since it wouldn't matter anyway. When the round's distribution was below \$5, I'd usually put about \$2 I'd be willing to pay, depending slightly on the percentages.
34	0	low	I stuck with relatively low values the entire time
35	0	spike	I either submitted the price that the item was most likely to be or a price closest to the most probable price.
36	1	[N/A]	[Failed to submit the end-of-experiment survey.]
37	1	low	for the first part i went about as low as i could,
38	0	bimodal	if 0 had a probability then I chose it, but if not I chose below the first dollar amount with probability
39	0	zero	I did not want the item
40	1	median	i saw the trends on the graphs and took my value to be a median of the numbers on the chart
41	1	zero	I didn't want to spend anything on cookies. Cookies are bad for you anyways.
42	0	random	Looked at probabilities. Since they were all fairly low, I tended to choose values randomly.
43	0	satisfaction	looked at the probabilities and decided betting extremely high would most likely allow me to get equal to if not higher than the actual number
44	1	predetermined	I submitted the value I placed on the item for each round, which was always 3.
45	1	predetermined	I had about the same answer for every round (\$3.00); the possibility of obtaining the item didn't matter because that's always the amount I'd pay (or slightly more, if the chances were more promising a slight but more significant percentage higher).
46	0	predetermined	0.50 was the most I was willing to pay for a dozen cookies.
47	0	predetermined	I valued the item at \$4
48	1	preference	Personal preference

Chapter 2. Testing Distributional Dependence in the Becker-DeGroot-Marschak Mechanism

Id	Det.	Category	Comment
49	1	other	instinct.
50	0	zero	as the round pass, I decided to not get the cookie since it seemed like I had to pay lot of money for it
51	0	median	I stayed around the middle of the range of numbers and guessed slightly on the lower side. If there was a higher probability, I guessed closer to that side.
52	1	predetermined	By what the item was worth to me.
53	1	predetermined	Basically according to my personal interest in the cookie, I don't want to put more money than \$1 on it.
54	0	probability	prob.
55	0	probability	i decided base on what percentages that im likely to get the item
56	1	predetermined	50 cents per cookie, \$6 total, for every round
57	1	probability	i chose amount to put in based on where the blue graph lines were. if there were many blue lines on the right side of the graph, extending very high, i would put a larger value. if there were higher blue bars on the left, my value would be where the graph lines stopped showing a high percentage. i didnt really keep the value of the cream certificate in mind, just the likelihood of whatever amount i might put in that round would be mean i had a better chance at winning.
58	0	other	i thought that too high would decrease my money (and if i even got the cookies, id just binge), and too low would waste the chance at a free gift card
59	0	bimodal	If the distribute was over 5.00, I chose 4.00 so I would not be paying more for a card that I did not value over 3.00. If the distribution had over a 50% chance of being under 3.00, I chose 2.50 for my value.
60	1	hybrid	by how likely it was that I could get it for a specific price
61	1	bimodal	If there was an equal amount of chance for the item to be at a low cost, then I would bet a low amount of money, but if it was a graph with a probability that increased to a higher number, I would not bet any money, and if there was no probability for anything lower than a dollar I would not bet any mother.
62	0	hybrid	I did not want to pay more than 5. If it was very likely to be less than five my willingness to pay went dow[n] because i expected it to be less than 5.
63	0	predetermined	I put \$5-6 on all rounds as that is what I am willing to sacrifice for it. And I would be perfectly content if in 1 round the value of it was rolled as 0 and I gave \$5-6 for something I could have gotten for 0.
64	1	predetermined	I looked at the charts, but basically I had already decided how much I wanted to risk.
65	1	preference	Based on personal preference.
66	0	mean	by thinking about the average values.
67	0	4 th quartile	Whichever value was the highest, I would enter that value
68	1	bimodal	If there was a high likelihood for the price to around 3.00 or less, I valued it at 3.50. If it was unlikely, I kept the price at 3.00.
69	1	probability	Based on probability of values

Chapter 3

Broken or Fixed Effects?

with Charles E. Gibbons

Department of Economics, University of California, Berkeley

and Juan Carlos Suárez Serrato

Stanford Institute for Economic Policy Research

3.1 Introduction

Fixed effects are a common means to “control for” unobservable differences related to particular qualities of the observations under investigation; examples include age, year, or location in cross-sectional studies or individual or firm effects in panel data. While fixed effects permit different mean outcomes between groups conditional upon covariates, the estimates of treatment effects are required to be the same; in more colloquial terms, the intercepts of the conditional expectations may differ, but not the slopes. An established result is that fixed effects regressions average the group-specific slopes proportional to both the conditional variance of treatment and the proportion of the sample in each group.¹ Researchers may believe that assuming a fixed effects model provides a convenient approximation of the sample-weighted effect and that models that incorporate group-specific effects yield estimates with significantly larger variances. In contrast to these beliefs, our replications of nine influential papers

¹See, e.g., Angrist and Krueger (1999); Wooldridge (2005a); Angrist and Pischke (2009).

reveals large differences between these estimates without large increase in variances.

This chapter empirically demonstrates large differences between the estimate from a fixed effects model and an average of treatment effects weighted only by the sample frequency of each group, our desired estimand. To identify this parameter, we interact the treatment variable with the fixed effects to identify a separate effect for each group and to average these estimates weighted by sample frequencies. Our approach can be applied to a broad array of questions in applied microeconomics. We demonstrate the generality of our point by examining nine papers from the *American Economic Review* between 2004 and 2009.² We choose these papers because they are among the most highly cited articles from this period in the *AER* and are widely considered as important pieces in their fields.³

The replication exercise demonstrates that, across a variety of units and groups of analysis, there are economically and statistically significant differences between the fixed effect estimate and the sample-weighted estimate. We employ the specification test that we develop to show that 6 of the 9 papers that we consider have sample-weighted estimates that are statistically different from the standard fixed effects estimates. Additionally, 7 of the 9 papers have estimates that differ in an economically significant way (taken here to mean differences of at least 10%). Averaging the largest deviance for each paper gives over 50% difference in the estimated treatment effect. We also show that our procedure does not markedly increase the variance of the estimator in 7 of 9 papers. While some of these papers do include interactions or run separate regressions for different groups, we show that there may be other statistically and substantively important interactions that might offer more informative estimates.

Our chapter begins by situating our approach in the literature in Section 3.2. In Section 3.3, we precisely define the parameter of interest in the presence of het-

²For a discussion of how these papers were chosen, see Appendix 3.C.1. An earlier draft of this chapter had a stronger emphasis on the returns to education literature and included an analysis of the results of Acemoglu and Angrist (2000).

³Thanks to a recent policy decision by the editorial board of the *AER*, it is possible to access the data and programs used in recently published articles and to replicate the results of these studies. We only analyze the data that the authors provide openly on the EconLit website. Though some of these papers include both OLS and instrumental variables approaches, we consider the implications of our approach for the OLS specifications to focus on the weighting scheme applied in this procedure.

erogeneity and show that FE models in this context are inconsistent estimators for the sample-weighted average except in special cases. We derive a test that distinguishes between the sample-weighted average and the FE estimate. To illustrate these results through an empirical example, in Section 3.4, we use a simplified model from Karlan and Zinman (2008) to compare the weighting scheme from the FE model to a sample-weighted approach and study the implications for the final estimate. We demonstrate the generality of these points in Section 3.5 in which we replicate eight other influential papers. We conclude in Section 3.6 by offering guidance to the applied researcher.

3.2 Incorporating heterogeneous treatment effects

In the presence of heterogeneous treatment effects across groups in the sample, the FE estimator gives an average of these effects. These weights depend not only on the frequency of the groups, but also upon sample variances within the groups. Angrist and Krueger (1999) compare the results from regression and matching estimators, demonstrating that the effects of a dichotomous treatment are averaged using different weights in each procedure.⁴ Closest to our derivation below, Wooldridge (2005*a*) finds sufficient conditions for FE models to produce sample-weighted averages in correlated random coefficient models. Our analysis builds upon this derivation for the case of fixed coefficients and offers a different interpretation of the necessary conditions for this result. Additionally, while these papers provide a strong theoretical reason to believe that FE estimators do not provide sample-weighted estimates, we illustrate the empirical importance of this distinction using a broad array of microeconomic questions.

There has long been an interest in coefficient heterogeneity across cross-sectional groups. A notable early piece is Chow (1960). Here, he runs regressions separately by group, which is the most flexible way of permitting heterogeneity across these groups for a given model, and compares the predictive power of the separate regressions to that of the pooled regression, forming a test for differences in slopes and intercepts. We begin with a test in the same spirit, but we only test for different treatment effects and use a test robust to heteroskedasticity by using a Wald test. Our suggested means

⁴See also Angrist and Pischke (2009).

of incorporating heterogeneous treatment effects is through interaction terms, a less flexible, but more parsimonious solution.

Many studies, including many of those that we replicate in this chapter, run separate regressions by group precisely because of the presence of treatment effect heterogeneity. Less common is the interacted model that we propose. Notable exceptions include Heckman and Hotz (1989), who consider the specific case of individual-specific time trends, which they call the random growth rate model. Papke (1994) and Friedberg (1998) also use the random growth model and find that the results of their studies are greatly influenced by trends that vary across geographic districts.

These examples, however, use interactions on predictors to avert omitted variables bias or to improve the fit of their models. In a different approach, Lochner and Moretti (2011) consider non-linearities in treatment effects, but do not estimate heterogeneous treatment effects across groups as we do here. In contrast to these works, the point of our analysis is that models that do not account for heterogeneous effects may provide inconsistent estimates of average effects.

We extend this literature in three ways. First, while Wooldridge (2005*a*) gives the sufficient conditions for a fixed effects model to deliver the sample-weighted treatment effect, we offer an alternative exposition and show what estimate is given by a FE model when this assumption fails. We focus on treatment effect heterogeneity and illustrate how it can be characterized and incorporated into a model in a parsimonious manner. Next, we derive a test that can distinguish between sample-weighted estimates derived from an interacted model and FE estimates. Our most important contribution is to show that these models are broadly empirically relevant in the the applied economics literature.

3.3 Interpreting FE estimates using projection results

In this section, we consider a specific model of heterogeneous treatment effects. Intuition might lead us to believe that, in the presence of heterogeneous treatment effects, FE estimates are sample-weighted averages of the group-level effects, the implicit parameter of interest. Instead, it has been established that, though the estimates are weighted combinations of group effects, they are not weighted by the size of the

group; instead, these weights depend upon sample variances. We illustrate this point by applying the Frisch-Waugh-Lovell theorem to the fixed effects model.

3.3.1 FE model estimates compared to the SWE

Suppose that a researcher estimates a fixed-effects model using data arising from a process with heterogeneous treatment effects given by

$$\begin{aligned} y_{ig} &= \alpha_g + \mathbf{w}_i\boldsymbol{\gamma} + x_i\beta_g + \nu_i \\ &= \alpha + (\alpha_g - \alpha)\mathbb{I}_g + \mathbf{w}_i\boldsymbol{\gamma} + x_i\beta + x_i\mathbb{I}_g(\beta_g - \beta) + \nu_i \\ \mathbf{y} &= \mathbf{Z}_{INT}\boldsymbol{\theta}_{INT} + \boldsymbol{\nu}, \end{aligned} \tag{3.1}$$

where the effect of interest, β_g , is group-specific. In this model, x_i is treatment, \mathbb{I}_g is a vector of group fixed effects, and \mathbf{w}_i is a vector of additional covariates.⁵ Though it may be instructive to consider the heterogeneity in these effects across groups, researchers often want a single summary of the treatment effect. A natural candidate would be the sample-weighted treatment effect, as explored in Wooldridge (2005b), as an example.

Definition 1 (Sample-weighted treatment effect). *The sample-weighted treatment effect for the model in Equation 3.1 is*

$$\bar{\beta} = \sum_g \widehat{\Pr}(g)\beta_g,$$

where $\widehat{\Pr}(g) = \frac{N_g}{N}$, N is the total number of observations in the sample and N_g is the number of observations belonging to fixed effect group $g \in 1, \dots, G$.

Definition 2 (Sample-weighted coefficient estimates). *The sample-weighted coefficient estimates from an interacted model with regression coefficients $\widehat{\boldsymbol{\theta}}_{INT}$ are*

$$\widehat{\boldsymbol{\theta}}_{SWE} = \mathbf{W}\widehat{\boldsymbol{\theta}}_{INT} \equiv [\mathbf{I}_K \mathbf{F}_0]\widehat{\boldsymbol{\theta}}_{INT},$$

⁵Though there are G groups, there are $G - 1$ fixed effects included in the model for identification purposes. Assume that group G is the excluded group.

where \mathbf{I}_K is a K -dimensional identity matrix, with K being the number of covariates not involving treatment, and

$$\mathbf{F}_0 = \frac{1}{N} \begin{bmatrix} 0 & \dots & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & \dots & 0 \\ N_1 & N_2 & \dots & N_{G-1} \end{bmatrix}.$$

Suppose that the researcher estimates a FE model that contains a single treatment effect parameter,

$$\begin{aligned} y_{ig} &= a_g + \mathbf{w}_i \mathbf{c} + x_i b + u_i \\ \mathbf{y} &= \mathbf{A}_{FE} \boldsymbol{\theta}_{FE} + \mathbf{x} b + \mathbf{u}; \end{aligned}$$

here, \mathbf{A}_{FE} contains the fixed effects and covariates other than treatment. Following the Frisch-Waugh-Lovell theorem, we can find the coefficient estimate \hat{b} by multiplying both sides of this expression by the annihilator matrix $\mathbf{M}_A = I - (\mathbf{A}'\mathbf{A})^{-1} \mathbf{A}'$, giving

$$\begin{aligned} \mathbf{M}_A \mathbf{y} &= \mathbf{M}_A \mathbf{x} b + \mathbf{M}_A \mathbf{u} \\ \Rightarrow \hat{b} &= (\mathbf{x}' \mathbf{M}_A \mathbf{x})^{-1} \mathbf{x}' \mathbf{M}_A \mathbf{y} = \frac{\widehat{Cov}(\tilde{x}_i, y)}{\widehat{Var}(\tilde{x}_i)}, \end{aligned}$$

where \tilde{x}_i is the projected value of treatment for observation i .

The FE model above posits that the effect of treatment across groups is homogeneous. The OLS estimator \hat{b} is a consistent estimator of the sample-weighted effect only in special cases. Instead of a sample-weighted estimate, the FE estimator gives

$$\sum_{g \in G} \widehat{\Pr}(g) \hat{\beta}_g \left(\frac{\widehat{Var}(\tilde{x}_i | g)}{\widehat{Var}(\tilde{x}_i)} \right), \quad (3.2)$$

See Appendix 3.A.1 for a derivation of this result. We see that the FE and SWE are the same when the treatment effects are homogeneous or the variance of the projected treatment is the same across all groups. Otherwise, the FE estimator overweights groups that have larger variance of treatment conditional upon other covariates and

underweights groups with smaller conditional variances.

From Equation 3.2, we see that, while FE models do provide a weighted combination of group effects, these effects are not weighted by sample frequencies. Instead, these weights depend upon sample variances, thereby producing estimates that are less informative for policy analysis. The weighting scheme employed by FE models provides a more efficient estimate of the treatment effect *in the absence of heterogeneous treatment effects*. In the presence of heterogeneity, however, it does not produce an estimate that is readily interpretable or comparable across studies.

If the FE model is the true data-generating process, then there are homogeneous treatment effects. Hence, estimates arising from a an analysis using only subgroup of our sample should be identical to those obtained by examining the entire sample with fixed effects included. This implies that the estimate of the treatment effect is invariant to the distribution of the groups in the sample. If the FE model does not hold, then the FE estimate \hat{b} is a function of the sample covariances; this statistic may change across samples or in subsamples. As a result, estimates are sample-dependent and not comparable across subsamples or studies.

Proposition 4 (Sufficient condition for consistent estimation of sample-weighted treatment effects). *The fixed effects model consistently estimates the sample-weighted average in the presence of heterogeneous treatment effects if the variance of treatment conditional on all other covariates is the same across all groups; i.e. $\widehat{\text{Var}}(\tilde{x}_i | g) = \widehat{\text{Var}}(\tilde{x}_i) \forall g$. (see Appendix 3.A.1).*

Thus, a regression on data from a perfectly randomized experiment where treatment has the same variance across groups yields the sample-weighted treatment effect. Such perfection is likely unattainable in observational or experimental settings, however. Indeed, in Section 3.5, we replicate a randomized experiment in Karlan and Zinman (2008). In that experiment, treatment (an interest rate on a microloan in South Africa) is randomized within different fixed effects groups (the risk category of the borrower), but the ranges of the (multi-valued) treatment are not the same across groups and, as a result, neither are the variances. In this case, we find that the sample-weighted treatment effect differs from the FE estimate by 61%. We use this case study to quantitatively illustrate the proposition above in Section 3.4.

3.3.2 A Test of Equality Between Sample-Weighted and FE Estimates

Even if the included interactions are statistically significant, it could be that their sample-weighted average is not statistically different from the standard FE model that excludes these interactions. We derive a specification test to discriminate between the FE estimate and the sample-weighted average.

Proposition 5 (Specification Test of the differences between the FE estimates and the sample-weighted average). *The test of the following null hypothesis*

$$H_0 : \text{plim} \left(\hat{\boldsymbol{\theta}}_{SWE} - \hat{\boldsymbol{\theta}}_{FE} \right) = \mathbf{0}$$

$$H_a : \text{plim} \left(\hat{\boldsymbol{\theta}}_{SWE} - \hat{\boldsymbol{\theta}}_{FE} \right) \neq \mathbf{0},$$

can be conducted by noting that the Wald test statistic

$$H = \left(\hat{\boldsymbol{\theta}}_{SWE} - \hat{\boldsymbol{\theta}}_{FE} \right)' \left(N^{-1} \widehat{\text{Var}} \left[\hat{\boldsymbol{\theta}}_{SWE} - \hat{\boldsymbol{\theta}}_{FE} \right] \right)^{-1} \left(\hat{\boldsymbol{\theta}}_{SWE} - \hat{\boldsymbol{\theta}}_{FE} \right)$$

has an asymptotic $\chi^2(q)$ distribution under H_0 , where $q = \text{rank} \left(\hat{\boldsymbol{\theta}}_{SWE} - \hat{\boldsymbol{\theta}}_{FE} \right)$; H_0 is rejected at level α when $H > \chi^2_\alpha(q)$. Robust estimation of this test statistic is addressed in Appendix 3.A.2. This test is implemented by the Stata command `GSSUtest` discussed in Appendix 3.B.

This test compares all coefficients in both models. Other tests can also be conducted using $\left(\hat{\boldsymbol{\theta}}_{SWE} - \hat{\boldsymbol{\theta}}_{FE} \right)$ and $\widehat{\text{Var}} \left[\hat{\boldsymbol{\theta}}_{SWE} - \hat{\boldsymbol{\theta}}_{FE} \right]$ by imposing the necessary restrictions on H . For example, we provide t tests of the single null hypothesis that the estimate of the treatment effect from a FE model differs from the sample-weighted average in our meta-study in Section 3.5.

3.4 A Case Study: Karlan and Zinman (2008)

In this section, we provide a detailed case study of one of our selected *AER* papers. This example illuminates the exposition of Section 3.3.1 and further clarifies the relationship between the FE and sample-weighted estimates.

We show in Section 3.3.1 that if an experiment is perfectly randomized, then the FE estimate should equal the sample-weighted average. More specifically, all covariates need to be precisely uncorrelated with treatment within each group and the variance of treatment must be the same across all groups (see Equation 3.2). Among our *AER* replications, we have one experiment that we can consider more closely. Karlan and Zinman (2008) randomized the interest rate offered for a microloan across a population of South Africans. They look to identify the credit elasticity among this group.

In the case of Karlan and Zinman (2008), the authors include two sets of covariates other than the treatment: the financial risk of the borrower and the mailer wave of the experiment when the borrower participated. The distributions of treatment and risk level are nearly uncorrelated with the mailer wave, hence, we ignore these fixed effects in this section only for expository purposes. But, to offer interest rates commensurate with prevailing market rates, the authors needed to charge higher rates to higher risk individuals. Recall that differing means in treatment do not drive the difference between the FE and SWE estimates, but rather differences in variances.

The authors offer not only higher rates to higher risk borrowers, but also offer a greater range of rates to this group; the variance of treatment differs across the groups.⁶ As a result, the FE estimate will not be equal to the SWE if the responsiveness to interest rates varies across risk groups.

The FE weights are given in column 2 of Table 3.1. These are the variances of treatment by group multiplied by the sample frequency of that group. Using these weights and the group effect estimated from an interacted model, given in column 4 of Table 3.1, we can calculate the FE estimate; this estimate is given in the bottom row of the table.

We can compare the weights from a FE model to the sample frequencies used to calculate a sample-weighted average; these weights appear in column 3 of Table 3.1. We see that high risk individuals are overweighted in the FE model and the low and medium risk individuals are underweighted. This accords with the design of the

⁶Again, we assume that mailer wave is uncorrelated with treatment and drop it from the model that the authors actually employ. This is a reasonable assumption for these data. Hence, the variance conditional upon all covariates is just the variance of treatment by group.

study—high risk borrowers had a wider range of interest rate offers and this relatively high variance in treatment leads to overweighing in the FE estimate.

Differences in weighting scheme are only important if the treatment effect is heterogeneous. We find that high-risk borrowers are much less responsive to the interest rate than low-risk borrowers. Because high-risk individuals are overweighted and have a smaller (in absolute value) treatment effect, the FE estimate underestimates the responsiveness of individuals to the interest rate by nearly 70%.⁷

Table 3.1: Karlan and Zinman (2008) treatment effect weighting

Risk group	FE weight	Sample freq.	Effect
Low	0.045	0.125	-32.4
Medium	0.061	0.092	-9.9
High	0.894	0.783	-2.7
Average	-4.450	-7.050	

Notes: Note that the FE analogue here, -4.450, does not precisely equal the actual FE estimate of -4.37 due to correlation between mailer wave fixed effects and the interest rate (e.g. treatment).

3.5 Fixed Effects Interactions: An AER Investigation

We have seen that, even in randomized experiments, FE models generally do not provide the sample-weighted estimate in the presence of heterogeneous treatment effects. To produce the SWE, we propose using an interacted model, following Equation 3.1, where the treatment effects are summarized by averaging the interacted effects weighted by the sample frequency of each group. To examine the differences between FE models and our approach more broadly, we turn to highly cited papers published in the *American Economic Review* between 2004 and 2009. We choose this publication due to its influence and the quality of its papers and consider recent years in order to capitalize upon the *AER* editorial board’s decision to require posting of data and other replication details to the EconLit online repository. The papers that

⁷The estimate that we calculate is not precisely equal to the FE estimate given in the paper. This is because we did not include the mailer wave fixed effects, explaining the difference between cited differences of 61 and 70%.

we choose are well known in their respective fields and serve as prime examples of respected empirical work.

We find the nine most cited papers that use fixed effects in an OLS model as part of their primary specification and meet additional requirements, which serve to limit our scope to papers in applied microeconomics with a clear effect of interest. These papers are listed in Table 3.2 along with the outcomes, effects of interest, and fixed effects considered here. A complete description of the process that we follow to identify these papers can be found in Appendix 3.C.1 and a more detailed description of the regressions that we consider is given in Appendix 3.C.2.

3.5.1 Replication Results

To consider the importance of interactions in these papers, we first test the joint significance of the coefficients on the interactions between the effect of interest and the fixed effects using a standard Wald test. Then, we test whether a sample-weighted average arising from the interacted model differs from the estimate of the FE model.⁸ We develop a command called `GSSUtest` to perform these tests in Stata.⁹

Our results appear in Tables 3.3 and 3.4. This table provides the p -values for Wald tests of joint significance of the interaction terms and the single test of the difference between the sample-weighted treatment effect and the fixed effect estimate and the percent difference between the treatment effects. Additional detail is provided in Tables 3.7 through 3.14.

Every paper that we consider has at least one set of fixed effects interactions that is significant at the 5% level. Some authors correctly separate regressions to account for these issues. For example, Lochner and Moretti (2004) are correct in separating their regressions by race, an alternative to adding interaction terms. Card, Dobkin and Maestas (2008) are the most aggressive in the use of separate regressions, dividing the sample into education-by-race categories; the results suggest that this is merited. The use of separate regressions and interaction terms by all the authors is detailed in Table 3.6. For most papers, there is a need to include fixed effects interactions in

⁸See Appendix 3.A.2 for details on this test.

⁹See Appendix 3.B. The authors have posted a copy of this code online for researchers interested in implementing this test.

the analysis and we recommend that authors explore this possibility.

Having demonstrated that fixed effects interactions are important covariates in these models using joint Wald tests, we now demonstrate that their inclusion produces sample-weighted averages that are statistically and economically different from estimates arising from the standard FE model. We define economically significant as a difference between the two estimates of more than ten percent of the FE estimate.

Seven of these papers have differences that are economically significant, exceeding ten percent upwards to over three hundred percent; averaging the largest difference for each paper gives over a 50% difference in the estimated treatment effect with a median of 19.5%. Six of the nine papers have a set of interactions that produce a sample-weighted average that is individually statistically different from the FE estimate at the 5% level.¹⁰ We note that our ability to distinguish between these two estimates is related to the power of the original analysis. These results are similar to those found by Graham and Powell (2010) in their case study on heterogeneous treatment effects. It is crucial that policy makers calibrate the estimates that they obtain from the sample to their population of interest in order to obtain accurate and informative economic assessments. Fixed effects interactions provide a way of obtaining estimates relevant for policy analysis.

3.5.2 The interacted and FE models and the variance-bias trade-off

Our implementation of the interacted model incorporates group-specific treatment effects into a standard fixed effects regression. The choice between the standard FE model and the interacted version, then, can be viewed as the choice between short and long versions of a regression. The preceding discussion focuses on the bias of FE estimators relative to the SWE in a world of treatment effect heterogeneity. But, we are concerned with the variance of our estimators as well.

¹⁰We may be worried about multiple testing issues here. A conservative Bonferroni correction states that, for a set of n hypotheses, we can reject the joint null that all n null hypotheses are true with size α if we can reject any hypothesis individually at the $\frac{\alpha}{n}$ level. Since we obtain p -values on the order of 0.000, we can reject the joint null that all the sample-weighted averages equal the FE estimates.

Suppose that the variance of our estimates is lower in the FE model relative to the interacted model. Goldberger (1991) provides rationales for short, potentially biased, regression over a long regression that has higher variance using the variance-bias tradeoff framework. We consider these rationales in the context of FE and interacted models using the empirical evidence found in our meta-study. They are:

- The researcher believes that $\theta_{INT,2} = 0$; e.g. treatment effects are homogeneous and thus the coefficients on the interactions are expected to be zero. Fortunately, this assumption can be tested using a joint significance test of the coefficients on the interaction variables. These interactions are significant in a vast majority of the cases that we consider, rendering this an inappropriate justification for choosing the FE model.
- The researcher believes that $\theta_{INT,2} \neq 0$, but might accept an imperfect approximation θ_{FE} with smaller standard errors. This choice depends upon the magnitude of the difference between the estimators. We find that the difference between the FE estimate and a sample-weighted average exceeds 10% in eight of the nine papers that we consider and averaging the largest deviations from each paper gives a difference of 50% between the treatment effects; the difference between the estimators is often substantial and consequential for policy analysis.

To evaluate the variance-bias tradeoff in our replications, we can examine the relationship between the largest absolute difference for each paper and compare that to the percent difference in standard error of the treatment effect between the two models; Figure 3.1 shows this relationship.¹¹ We see that, for seven of the papers, the variance does not substantively increase when calculating the SWE from an interacted model; indeed, it decreases for four of these papers. Hence, for these papers, it is not necessary to accept an imperfect estimate in order to achieve reduced standard errors.

¹¹If the difference in the standard errors is positive, the SWE from the interacted model has a larger standard error. For Griffith, Harrison and Van Reenen (2006), the absolute difference is 324% and the percent change in standard errors is 630%; we exclude this outlier from the plot.

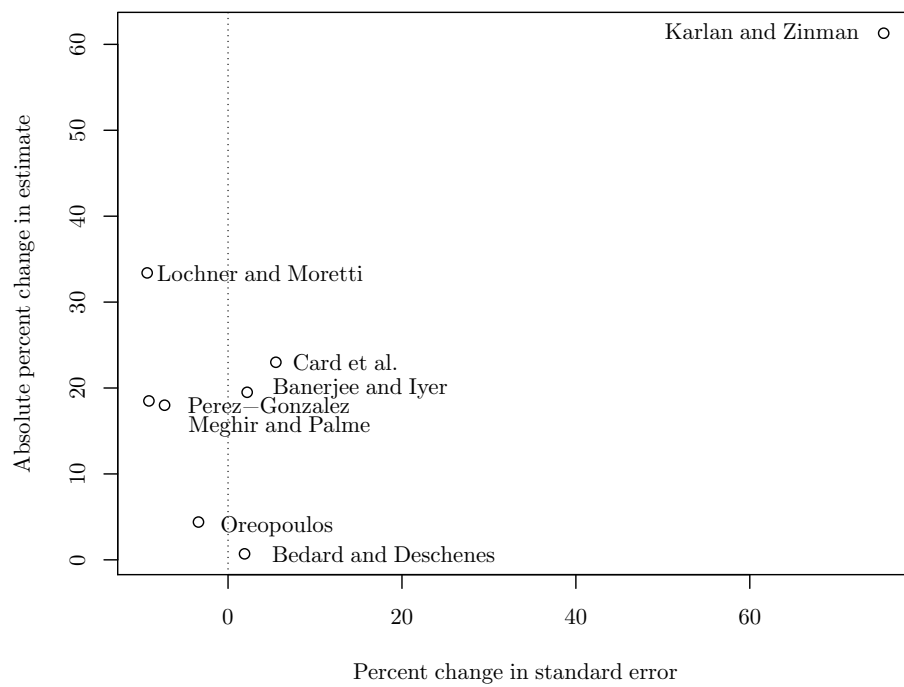


Figure 3.1: The relationship between the difference in the estimators and the change in variance among the *AER* replications

Table 3.2: Papers from the *AER* used in the meta-analysis

Citation	Outcome	Effect of interest	Fixed effects
Banerjee and Iyer (2005)	Fertilizer use	Prop. non-landlord land	Coastal dummy, year
	Proportion irrigated		
	Proportion other cereals		
	Proportion rice		
	Proportion wheat		
	Proportion white rice		
	Rice yield (log)		
	Wheat yield (log)		
Bedard and Deschênes (2006)	Smoking dummy	Veteran status	Age, education, race, region
Card et al. (2008)	Saw doctor dummy	Age over 65 dummy	Ethnicity, gender, region,
	Was hospitalized dummy		year, education level
Griffith et al. (2006)	Output-capital ratio (log)	U.S. Industry patents × U.S. presence	Industry, year
Karlan and Zinman (2008)	Loan size	Interest rate (log)	Mailer wave, risk category
Lochner and Moretti (2004)	Imprisonment	Education	Race, age, year
Meghir and Palme (2005)	Wage (log; change in)	Education reform	High ability dummy, high father's education dummy, sex, year
Oreopoulos (2006)	Wage (log)	Education	Age, Northern Ireland
Pérez-González (2006)	Market-book ratio	CEO heir inheritance	High family ownership
	Operating returns	dummy, year	

Notes: Additional details on our replications are found in Appendix 3.C.

Table 3.3: *AER* replication results

Citation	Fixed effect	Joint test	Diff. test	% diff.
Banerjee and Iyer (2005) (Prop. irrigated)	Coastal	0.231	0.827	-1.1
	Soil — black	0.387	0.482	4.7
	Soil — red	0.080	0.172	19.5†
	Soil — other	0.555	0.649	2.0
	Year	0.000**	0.901	0.0
Bedard and Deschênes (2006)	Age	0.944	0.914	0.1
	Education	0.002**	0.374	0.7
	Race	0.080	0.089	0.5
	Region	0.701	0.218	0.2
Card et al. (2008)	Ethnicity‡ (saw doctor)	0.000**	0.044*	1.3
	Gender	0.000**	0.665	0.8
	Region	0.156	0.882	-0.1
	Year	0.067	0.004**	-23.0†
	Education (whites)‡	0.004**	0.002**	-12.5†
	Education (non-whites)‡	0.771	0.323	-1.3
	Ethnicity (hospitalized)‡	0.000**	0.459	0.5
	Gender	0.000**	0.012*	-1.3
	Region	0.015*	0.732	0.2
	Year	0.778	0.722	0.3
	Education (whites)‡	0.003**	0.048*	1.4
	Education (non-whites)‡	0.746	0.295	5.7
Griffith et al. (2006)	Industry	0.000**	0.016*	-324.3†
	Year	0.040*	0.050*	6.5

Notes: Column 3 gives the p -value for the test of the joint significance of the interaction terms using a Wald test. Column 4 gives the p -value for a t test of the difference between the sample-weighted estimate and the FE estimate. Column 5 gives the percent difference between these two estimates. A single star indicates significance at the 5% level; two stars indicate significance at the 1% level. A dagger indicates a difference of more than 10% between the two estimates. A double dagger indicates whether the author considers heterogeneity among these groups. Results for two outcomes of interest are reported for Card et al. (2008); those outcomes are indicators for whether the individual saw a doctor or was hospitalized in the previous year.

Table 3.4: *AER* replication results, continued

Citation	Fixed effect	Joint test	Diff. test	% diff.
Karlan and Zinman (2008)	Mailer wave	0.330	0.837	-1.1
	Risk category	0.016*	0.010*	61.3†
Lochner and Moretti (2004)	Race‡ (all)	0.000**	0.000*	-0.9
	Age (blacks)	0.000**	0.000**	33.4†
	Year (blacks)	0.000**	0.000**	2.4
	Age (whites)	0.000**	0.000**	30.9†
	Year (whites)	0.002**	0.286**	0.22
Meghir and Palme (2005)	High father's education‡	0.000**	0.244	18.5†
	Sex‡	0.527	0.747	0.2
	Year	0.000**	0.013*	0.5
Oreopoulos (2006)	N.Ireland‡	0.000**	0.000**	4.4
	Age (GB)	0.000**	0.360	1.4
	Age (NI)	0.000**	0.150	-2.7
	Age (NI & GB)	0.000**	0.634	0.6
Pérez-González (2006)	Family ownership (MB)	0.223	0.243	18.0†
	Family ownership (OR)	0.483	0.489	10.4†
	Year (MB)	0.002**	0.329	-11.4†
	Year (OR)	0.010**	0.829	-2.4

Notes: Column 3 gives the p -value for the test of the joint significance of the interaction terms using a Wald test. Column 4 gives the p -value for a t test of the difference between the sample-weighted estimate and the FE estimate. Column 5 gives the percent difference between these two estimates. A single star indicates significance at the 5% level; two stars indicate significance at the 1% level. A dagger indicates a difference of more than 10% between the two estimates. A double dagger indicates whether the author considers heterogeneity among these groups.

3.6 Conclusion

This chapter contributes to the applied econometrics literature by illustrating a common issue in the application of fixed effects. Fixed effects are commonly employed to “control for” differences between groups. In the presence of heterogeneous treatment effects, researchers may intuitively believe that their estimates are sample-weighted averages of the group treatment effects. Though this is generally the parameter of interest, it is generally not the parameter that is identified by standard fixed effects models. We demonstrate this point using econometric theory and characterize its relevance to empirical applications.

Using an application of the Frisch-Waugh-Lovell theorem, we show that fixed effects models do not estimate the sample-weighted average treatment effect. We offer a sufficient condition for this difference to be 0 asymptotically and give an intuitive explanation of what is estimated if this condition is not met. We provide statistical tools to assess the importance of interaction terms, including a statistical test for the difference between the fixed effects estimate and the sample-weighted average from an interacted model. By employing these techniques, researchers can find estimates that are easier to interpret, that can be compared across academic studies, and that are more relevant for policy analysis.

While the sample-weighted average may be the most informative single statistic of the treatment effect for a sample, even it may not be the most relevant result for policy analysis. By identifying different effects for each subgroup, researchers can characterize patterns of treatment effect heterogeneity, permitting them to conduct more appropriate policy analysis and produce results that are comparable across academic studies. This process also generates a more flexible functional form that can better approximate the true data generating process.

Results from a replication exercise show that fixed effects interactions are significant in every paper that we consider across a variety of effects of interest and outcomes. The sample-weighted estimate is statistically different from the fixed effects estimate in six papers of the nine papers that we consider and substantively different in seven; using the largest difference for each paper, the average difference across replications is over 50%. Our results also show that we can achieve our de-

sired estimand without accepting an increase in variance. Finally, while authors often include interactions or run regressions separately for different subpopulations, incorporating these heterogeneous effects into a meaningful summary of mean effects would provide a better characterization of the data generating process without a substantial increase in variance.

Appendix 3.A Topics in Fixed Effects Theory

3.A.1 Sufficient Conditions for Estimation of Sample-Weighted Treatment Effects in FE Models

Suppose that a researcher estimates a fixed-effects model

$$\begin{aligned} y_{ig} &= \alpha_g + \mathbf{w}_i\boldsymbol{\gamma} + x_i b + e_i \\ &\equiv \mathbf{a}_i\boldsymbol{\delta} + x_i b + e_i, \end{aligned}$$

where \mathbf{a}_i contains the fixed effects and covariates other than treatment, x_i . Stacking these equations across all observations i gives

$$\mathbf{y} = \mathbf{A}\boldsymbol{\delta} + \mathbf{x}b + \mathbf{e}.$$

Following the Frisch-Waugh-Lovell theorem, we can find the coefficient b by multiplying both sides of this expression by the annihilator matrix $\mathbf{M}_A = I - (\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'$, giving

$$\begin{aligned} \mathbf{M}_A\mathbf{y} &= \mathbf{M}_A\mathbf{x}b + \mathbf{M}_A\mathbf{u} \\ \Rightarrow \hat{b} &= (\mathbf{x}'\mathbf{M}_A\mathbf{x})^{-1}\mathbf{x}'\mathbf{M}_A\mathbf{y} = \frac{\widehat{Cov}(\tilde{x}_i, y)}{\widehat{Var}(\tilde{x}_i)}, \end{aligned}$$

where \tilde{x}_i is the projected value of treatment for observation i . Define the group-specific effect as

$$\hat{\beta}_g = \frac{\widehat{Cov}(\tilde{x}_i, y | g)}{\widehat{Var}(\tilde{x}_i | g)}.$$

We can decompose the estimate of \hat{b} following

$$\begin{aligned}\hat{b} &= \frac{\widehat{Cov}(\tilde{x}_i, y_i)}{\widehat{Var}(\tilde{x}_i)} \\ &= \frac{\sum_{g \in G} \widehat{\Pr}(g) \widehat{Cov}(\tilde{x}_i, y_i | g)}{\widehat{Var}(\tilde{x}_i)} \\ &= \frac{\sum_{g \in G} \widehat{\Pr}(g) \hat{\beta}_g \widehat{Var}(\tilde{x}_i | g)}{\widehat{Var}(\tilde{x}_i)} \\ &= \sum_{g \in G} \widehat{\Pr}(g) \hat{\beta}_g \left(\frac{\widehat{Var}(\tilde{x}_i | g)}{\widehat{Var}(\tilde{x}_i)} \right).\end{aligned}$$

The second equality follows because we are considering a specific type of covariate—binary fixed effects. Thus, it is clear that the estimate of the treatment effect arising from the fixed effects model is not simply a frequency-weighted average of the group-specific effects. This is only the case if the conditional variances of the treatment within each group are the same.

The bias of the FE model in estimating the sample-weighted average, $\bar{\beta}$, has the following limit:

$$\begin{aligned}\text{plim}(\bar{\beta} - \hat{b}) &= \sum_g \left(\Pr(\mathbb{I}_g = 1) - \frac{\Pr(\mathbb{I}_g = 1) \text{Var}(x | \mathbb{I}_g = 1)}{\text{Var}(x)} \right) \beta_g \\ &= \sum_g \Pr(\mathbb{I}_g = 1) \left(1 - \frac{\text{Var}(x | \mathbb{I}_g = 1)}{\text{Var}(x)} \right) \beta_g.\end{aligned}$$

Again, this difference is 0 if $\text{Var}(\tilde{x}_i | g) = \text{Var}(\tilde{x}_i)_i \forall g$.

3.A.2 Calculating the Difference Between the Fixed Effects and Weighted Interactions Estimators

We may wonder whether the difference between the FE model estimate of the treatment effect is statistically significantly different from a sample-weighted estimate of the treatment effect arising from the interacted model. Define the fixed-effects model

(FE) as

$$\begin{aligned} y_{ig} &= a_g + \mathbf{w}_i \mathbf{c} + \mathbf{x}_i \mathbf{b} + u_i \\ \mathbf{y} &= \mathbf{Z}_{FE} \boldsymbol{\theta}_{FE} + \mathbf{u} \end{aligned}$$

and the interacted model as

$$\begin{aligned} y_{ig} &= \alpha_g + \mathbf{w}_i \boldsymbol{\gamma} + \mathbf{x}_i \boldsymbol{\beta}_g + \nu_i \\ \mathbf{y} &= \mathbf{Z}_{INT} \boldsymbol{\theta}_{INT} + \boldsymbol{\nu}, \end{aligned}$$

where i indexes the individual unit from 1 to N , g indexes group membership from 1 to G , and $\boldsymbol{\theta}'_{FE} = [a_1, \dots, a_G, \mathbf{c}', \mathbf{b}']$, and $\boldsymbol{\theta}'_{INT} = [\alpha_1, \dots, \alpha_G, \boldsymbol{\gamma}', \boldsymbol{\beta}'_1, \dots, \boldsymbol{\beta}'_G]$. The crucial difference between these two models is that the interacted model allows the coefficient on \mathbf{x}_i to vary across groups.

The test that we propose considers whether the sample-weighted average of $\boldsymbol{\beta}_g$ in the interacted model equals \mathbf{b} from the FE model. We derive the distribution of the test statistic through joint estimation of the models using a Method of Moments (MM) approach. We first derive the joint distribution of the estimators, then we develop a specification test for our particular hypothesis.

For these models, the sets of moment conditions are given by:

$$\begin{aligned} \sum_{i=1}^N \mathbf{h}_{FE,i}(\widehat{\boldsymbol{\theta}}_{FE}) &\equiv \sum_{i=1}^N \mathbf{z}_{FE,i} (y_{ig} - \mathbf{z}_{FE,i} \widehat{\boldsymbol{\theta}}_{FE}) = \mathbf{0} \quad \text{and} \\ \sum_{i=1}^N \mathbf{h}_{INT,i}(\widehat{\boldsymbol{\theta}}_{INT}) &\equiv \sum_{i=1}^N \mathbf{z}_{INT,i} (y_{ig} - \mathbf{z}_{INT,i} \widehat{\boldsymbol{\theta}}_{INT}) = \mathbf{0}. \end{aligned}$$

Stacking these equations into $\sum_{i=1}^N \mathbf{h}_i(\widehat{\boldsymbol{\delta}}) = \mathbf{0}$, where $\widehat{\boldsymbol{\delta}}' = [\widehat{\boldsymbol{\theta}}'_{FE}, \widehat{\boldsymbol{\theta}}'_{INT}]$ and $\boldsymbol{\delta}'_0 = [\boldsymbol{\theta}'_{FE}, \boldsymbol{\theta}'_{INT}]$, and applying standard MM arguments (see, e.g. Cameron and Trivedi, 2005), it follows that $\widehat{\boldsymbol{\delta}}$ has the property that

$$\sqrt{N} (\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}_0) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{G}_0^{-1} \mathbf{S}_0 (\mathbf{G}'_0)^{-1}),$$

where

$$\mathbf{G}_0 = \text{plim} \frac{1}{N} \sum_{i=1}^N \left[\frac{\partial \mathbf{h}_i}{\partial \boldsymbol{\delta}'} \bigg|_{\boldsymbol{\delta}=\boldsymbol{\delta}_0} \right] \quad \text{and} \quad \mathbf{S}_0 = \text{plim} \frac{1}{N} \sum_{i=1}^N \sum_{j=1}^N \left[\mathbf{h}_i \mathbf{h}_j' \bigg|_{\boldsymbol{\delta}=\boldsymbol{\delta}_0} \right].$$

Note that, by partitioning the matrix $\mathbf{G}_0 = \begin{bmatrix} \mathbf{G}_{11} & \mathbf{G}_{12} \\ \mathbf{G}_{21} & \mathbf{G}_{22} \end{bmatrix}$ and using the fact that

$$\frac{\partial \mathbf{h}_{i,FE}}{\partial \boldsymbol{\theta}'_{INT}} = \mathbf{0} \quad \text{and} \quad \frac{\partial \mathbf{h}_{i,INT}}{\partial \boldsymbol{\theta}'_{FE}} = \mathbf{0},$$

it follows that $\mathbf{G}_{21} = \mathbf{G}_{12} = \mathbf{0}$.

As is standard (once again, see Cameron and Trivedi, 2005), we estimate \mathbf{G}_0 via

$$\widehat{\mathbf{G}} = \frac{1}{N} \sum_{i=1}^N \left[\frac{\partial \mathbf{h}_i}{\partial \boldsymbol{\delta}'} \bigg|_{\boldsymbol{\delta}=\widehat{\boldsymbol{\delta}}} \right].$$

To estimate \mathbf{S}_0 we consider two cases. First, assuming independence over i , an estimator robust to heteroskedasticity is

$$\widehat{\mathbf{S}}_R = \frac{1}{N} \sum_{i=1}^N \mathbf{h}_i(\widehat{\boldsymbol{\delta}}) \mathbf{h}_i(\widehat{\boldsymbol{\delta}})'$$

A second estimator that incorporates clustered errors is

$$\widehat{\mathbf{S}}_C = \frac{1}{N} \sum_{c=1}^C \sum_{i=1}^{N_c} \sum_{j=1}^{N_c} \mathbf{h}_{ic}(\widehat{\boldsymbol{\delta}}) \mathbf{h}_{jc}(\widehat{\boldsymbol{\delta}})'$$

Thus, robust and clustered estimators of the variance of $\widehat{\boldsymbol{\delta}}$ are $\widehat{Var}[\widehat{\boldsymbol{\delta}}] = \widehat{\mathbf{G}}^{-1} \widehat{\mathbf{S}}_e (\widehat{\mathbf{G}}')^{-1}$ for $e = R, C$ respectively.

Now we turn to the specific hypothesis that we would like to consider; namely, that the sample-weighted averages of the estimates from the interacted model are

equal to the FE estimates. Specifically, our hypothesis is

$$\begin{aligned} H_0 &: \text{plim} \left(\mathbf{W}\hat{\boldsymbol{\theta}}_{INT} - \hat{\boldsymbol{\theta}}_{FE} \right) = \mathbf{0} \\ H_a &: \text{plim} \left(\mathbf{W}\hat{\boldsymbol{\theta}}_{INT} - \hat{\boldsymbol{\theta}}_{FE} \right) \neq \mathbf{0}, \end{aligned}$$

where \mathbf{W} is defined as

$$\mathbf{W} \equiv \left[\mathbf{I}_Q, \begin{bmatrix} \mathbf{0}_{(Q-1 \times K-1)} \\ \mathbf{f} \end{bmatrix} \right]$$

to produce a sample-weighted estimate of the treatment effect and to return the other parameters.¹² In this formulation, Q is the rank of \mathbf{Z}_{FE} , \mathbf{I}_Q is a $Q \times Q$ identity matrix, K is the number of fixed-effect groups, and \mathbf{f} is a $[1 \times K - 1]$ vector of sample frequencies of fixed effect group membership.

To compute the difference of the estimators, define the matrix

$$\mathbf{R} = [-\mathbf{I}_Q, \mathbf{W}].$$

Then, the difference between the estimators is $\mathbf{R}\hat{\boldsymbol{\delta}} = \mathbf{W}\hat{\boldsymbol{\theta}}_{INT} - \hat{\boldsymbol{\theta}}_{FE}$ and the variance of this difference is estimated according to

$$\widehat{Var}[\mathbf{W}\hat{\boldsymbol{\theta}}_{INT} - \hat{\boldsymbol{\theta}}_{FE}] = \mathbf{R}\widehat{Var}[\hat{\boldsymbol{\delta}}]\mathbf{R}'.$$

The Wald test statistic

$$H = \left(\mathbf{W}\hat{\boldsymbol{\theta}}_{INT} - \hat{\boldsymbol{\theta}}_{FE} \right)' \left(N^{-1}\widehat{Var} \left[\mathbf{W}\hat{\boldsymbol{\theta}}_{INT} - \hat{\boldsymbol{\theta}}_{FE} \right] \right)^{-1} \left(\mathbf{W}\hat{\boldsymbol{\theta}}_{INT} - \hat{\boldsymbol{\theta}}_{FE} \right)$$

has an asymptotic $\chi^2(q)$ distribution under H_0 ; H_0 is rejected at level α when $H > \chi^2_\alpha(q)$. This test compares all coefficients in both models. Other tests can also be conducted using $\left(\mathbf{W}\hat{\boldsymbol{\theta}}_{INT} - \hat{\boldsymbol{\theta}}_{FE} \right)$ and $\widehat{Var} \left[\mathbf{W}\hat{\boldsymbol{\theta}}_{INT} - \hat{\boldsymbol{\theta}}_{FE} \right]$ by imposing the necessary restrictions on H . For example, we provide t tests of the single null hypothesis that the estimate of the treatment effect from a FE model differs from the

¹²Recall that, in our case, \mathbf{x}_i is a scalar.

sample-weighted average in our meta-study.

Appendix 3.B `GSSUtest.ado`

As a companion to this chapter, we develop a Stata command called `GSSUtest` that computes the sample-weighted average treatment effect, tests for equality of coefficients with those of a fixed effects model, and computes the percentage change in the parameter of interest. The command is available from the authors and can be executed with the following syntax:

```
GSSUtest y Tr FEg [varlist] [if] [in] [, options]
```

where

- `y` is the dependent variable,
- `Tr` is the independent variable of interest (e.g. treatment) and,
- `FEg` is a categorical variable indexing the fixed effect group.

Other predictors can be included in `varlist` and several options including sample weights and clustering are also available. `GSSUtest` automatically uses robust standard errors in its calculations.

Appendix 3.C AER Replications

3.C.1 Paper Selection

We aim to show the broad importance of these fixed effects interactions in capturing the sample-weighted average treatment effect. We do this by replicating high quality papers from a variety of fields. We begin by outlining guidelines for inclusion in our analysis:

- The paper must be in the *American Economic Review*. We enact this qualification in order to limit our universe of analysis both in terms of quantity and quality of papers and to guarantee easy access to the necessary data.

- The paper must be published in the March 2004 issue or later (to March 2009, the issue predating our literature search). The *AER* policy during this period requires that, barring any acceptable restriction, data for these papers be posted to the EconLit website. This leads to the condition that:
- The data necessary to replicate the main specification(s) of the paper must be readily available on the EconLit website.¹³ We use these data and direct those interested to the EconLit website to obtain these files.
- The main specification(s) of the paper must have a specific effect of interest.
- The main specification(s) of the paper must use some type of fixed effect. We identify papers meeting this qualification by searching the PDF files of the published papers for the terms “fixed effect” (which captures the plural “effects” as well) and for “dumm” (which captures “dummy” or “dummies,” common synonyms for fixed effects).
- We limit ourselves to microeconomic analyses and do not consider papers based on financial economics issues.
- We ignore papers that require special methods to incorporate time series issues.

We choose to replicate a total of nine papers in our analysis. To order our search, we consider papers in order of citations per year since publication. First, we use the citation counts provided by the ISI Web of Science on July 16, 2009. We limit our search to the *American Economic Review* and years 2004–2009, as outlined above. Unfortunately, the Web of Science does not provide the volume for the papers contained therein. We create an algorithm that assigns a volume number to a paper based upon its page number; these assignments are verified as papers are considered. The total number of citations are divided by the years since publication. For example, in June 2009, a paper published in June 2004 was published 5 years ago and a paper published in September 2004 was published 4.75 years ago.

¹³We determine which specifications are the “main” ones by considering the discussion of the effects in the text by the authors and ignore those specifications identified as robustness checks.

Citation counts are very noisy in the short time after publication that we consider here. Our citations-per-year metric might overweight later papers.¹⁴ Nonetheless, we consider all papers in this period with over 20 citations and 86% of all papers with 15 or more citations. It appears that we consider most of the highly cited papers from this period and do not ignore the most recent papers, as would occur using the gross citation count.

Papers that we select must be highly-cited and fit the qualifications necessary to be relevant to our inquiry; we replicate papers from 2004, 2005, 2006, and 2008, missing only 2007 and the one quarter of 2009 that predates our search. We examine a breadth of papers that covers several fields, several years, and several units of analysis and thus they serve as a decent representation of the use of fixed effects in the applied econometrics literature.

Before incorporating interaction terms into the specifications that we consider, we first ensure that we can replicate the results obtained by the authors as given in their respective papers. We can provide Stata `D0` and `log` files that generate and produce these results. We extend these files by incorporating the interactions as introduced in the paper. In choosing the interactions when there are several fixed effects in the regressions, we choose such that the number of groups is not unruly (U.S. states, for example, may simply produce too many terms to be informative). Our interacted regressions preserve all other features of the replicated specifications (e.g. clustering, robust standard errors, and inclusion of other covariates) unless otherwise noted in the text.

We do not justify that the interactions that we employ are the most salient within the given economic situation. Additionally, we do not suggest that the inclusion of interactions is the first-order extension of the analysis in the papers that we examine. We make no effort to search the subsequent literature to identify other areas of concern in these papers. Lastly, many of these papers employ instrumental variables to confront endogeneity. In these cases, we use the base OLS case to illustrate our point.

¹⁴In June 2009, 1 citation for a paper published in March 2009 is equal to 4 for a paper published in June 2008 and 20 for a paper published in June 2004.

3.C.2 Replication Details

We replicate the specifications cited in Table 3.5. Some of these authors include fixed effects interactions or run regressions separately for subgroups; we list these practices in Table 3.6. In Banerjee and Iyer (2005), the authors have eight separate outcomes of interest. In the body of the chapter, we give results only for a sample of these results. In Tables 3.11 through 3.14, we provide the results for all outcome-group combinations.

Table 3.5: Replication sources

Citation	Table	Column
Banerjee and Iyer (2005)	3	1
Bedard and Deschênes (2006)	5	1
Card, Dobkin and Maestas (2008)	3	6, 8
Griffith, Harrison and Van Reenen (2006)	3	2
Karlan and Zinman (2008)	4	1
Lochner and Moretti (2004)	3	1
Meghir and Palme (2005)	2	1 (row 1)
Oreopoulos (2006)	2	3
Pérez-González (2006)	9	1, 6

Notes: In Griffith, Harrison and Van Reenen (2006), we do not cluster at the industry level as the authors do in their paper. We also do not cluster as Oreopoulos (2006) does. In both cases, clustering does not change the results. We are not able to replicate the point estimate that Oreopoulos (2006) provides for his regression of Northern Ireland and Great Britain combined; we use the specification that he provides and base our results on this model.

Table 3.6: Fixed effects interactions and regressions by subgroup conducted in the original papers

Citation	Separate regressions	Interactions
Banerjee and Iyer (2005)	Entire country, subregion	
Bedard and Deschênes (2006)		
Card, Dobkin and Maestas (2008)	Race \times education	Age, age-squared
Griffith, Harrison and Van Reenen (2006)		
Karlan and Zinman (2008)		
Lochner and Moretti (2004)	Race (black, white)	
Meghir and Palme (2005)	Sex Father's education (low, high) Ability (low, high) Ability \times father's education \times sex	Sex
Oreopoulos (2006)	Country	
Pérez-González (2006)		Less selective college dummy Graduate school dummy Positive R&D spending dummy

Notes: Separate regressions and interaction terms only listed for specifications based upon the one given in Table 3.5. Pérez-González (2006) does not include the dummy variables that he subsequently interacts with treatment in his base regression; hence, we do not test their interactions here.

Table 3.7: Detailed replication results

Citation	Fixed effect	FE est.	FE SE	SWE	SWE SE
Bedard and Deschênes (2006)	Age	0.078	0.005	0.078	0.006
	Education	0.078	0.005	0.078	0.006
	Race	0.078	0.005	0.078	0.005
	Region	0.078	0.005	0.078	0.005
Card et al. (2008)	Ethnicity (saw doctor)	0.013	0.008	0.013	0.007
	Gender	0.013	0.008	0.013	0.008
	Region	0.013	0.008	0.013	0.008
	Year	0.013	0.008	0.010	0.008
	Education (whites)	0.006	0.008	0.006	0.008
	Education (non-whites)	0.039	0.013	0.039	0.014
	Ethnicity (hospitalized)	0.012	0.004	0.012	0.004
	Gender	0.012	0.004	0.012	0.004
Griffith et al. (2006)	Region	0.012	0.004	0.012	0.004
	Year	0.012	0.004	0.012	0.004
	Education (whites)	0.013	0.005	0.013	0.005
	Education (non-whites)	0.005	0.007	0.006	0.007
Griffith et al. (2006)	Industry	0.076	0.014	-0.170	0.104
	Year	0.076	0.014	0.080	0.014

Notes: Column 1 gives the paper and column 2 gives the fixed effects under consideration. Columns 3 and 4 give the standard FE model estimate of the treatment effect and its standard error. Columns 5 and 6 give the sample-weighted estimate from an interacted model and its standard error. Results for two outcomes of interest are reported for Card et al. (2008); those outcomes are indicators for whether the individual saw a doctor or was hospitalized in the previous year.

Table 3.8: Detailed replication results, continued

Citation	Fixed effect	FE est.	FE SE	SWE	SWE SE
Lochner and Moretti (2004)	Race (all)	-0.122	0.004	-0.121	0.003
	Age (blacks)	-0.370	0.015	-0.493	0.013
	Year (blacks)	-0.370	0.015	-0.379	0.015
	Age (whites)	-0.099	0.003	-0.130	0.002
	Year (whites)	-0.099	0.003	-0.099	0.003
Meghir and Palme (2005)	High father's ed.	0.014	0.009	0.017	0.008
	Sex	0.014	0.009	0.014	0.009
	Year	0.014	0.009	0.014	0.009
	N. Ireland	0.078	0.002	0.081	0.001
Oreopoulos (2006)	Age (GB)	0.075	0.002	0.076	0.002
	Age (NI)	0.106	0.004	0.104	0.003
	Age (NI & GB)	0.078	0.002	0.079	0.002
	High fam. own. (MB)	-0.256	0.086	-0.302	0.079
Pérez-González (2006)	High fam. own. (OR)	-0.027	0.009	-0.030	0.009
	Year (MB)	-0.256	0.086	-0.226	0.083
	Year (OR)	-0.027	0.009	-0.027	0.009
	Mailer wave	-4.368	1.093	-4.319	1.084
Karlan and Zinman (2008)	Risk category	-4.368	1.093	-7.047	1.917

Notes: Column 1 gives the paper and column 2 gives the fixed effects under consideration. Columns 3 and 4 give the standard FE model estimate of the treatment effect and its standard error. Columns 5 and 6 give the sample-weighted estimate from an interacted model and its standard error. The regression coefficients and standard errors from Lochner and Moretti (2004) are multiplied by 100, following the reporting of the authors in their paper.

Table 3.9: Detailed replication results, continued

Citation	Fixed effect	Joint test of interactions			Test of treat. diff.	
		Wald stat.	DF	<i>p</i> -value	<i>t</i> stat.	<i>p</i> -value
Bedard and Deschênes (2006)	Age	11.09	20	0.944	0.11	0.914
	Education	14.79	3	0.002	0.89	0.374
	Race	3.07	1	0.080	1.70	0.089
	Region	5.51	8	0.701	1.23	0.218
Card et al. (2008)	Ethnicity (saw doctor)	18.71	3	0.000	2.02	0.044
	Gender	114.37	1	0.000	0.43	0.665
	Region	5.23	3	0.156	-0.15	0.882
	Year	18.67	11	0.067	-2.88	0.004
	Education (whites)	13.13	3	0.004	-3.17	0.002
	Education (non-whites)	1.13	3	0.771	-0.99	0.323
	Ethnicity (hospitalized)	21.54	3	0.000	0.74	0.459
	Gender	18.50	1	0.000	-2.52	0.012
	Region	10.50	3	0.015	0.34	0.732
	Year	7.26	11	0.778	0.36	0.722
Griffith et al. (2006)	Education (whites)	13.99	3	0.003	1.98	0.048
	Education (non-whites)	1.23	3	0.746	1.05	0.295
	Industry	52.78	14	0.000	-2.40	0.016
	Year	19.04	10	0.040	1.96	0.050

Notes: Column 1 gives the paper and column 2 gives the fixed effects under consideration. Column 3 gives the Wald statistic of a joint test of the significance of the interactions, column 4 gives the degrees of freedom for that test, and column 5 gives the *p*-value. Column 6 gives a *t* statistic from a test of the difference between the FE and sample-weighted estimates using the test derived in Appendix 3.A.2 and the corresponding *p*-value. Results for two outcomes of interest are reported for Card et al. (2008); those outcomes are indicators for whether the individual saw a doctor or was hospitalized in the previous year.

Table 3.10: Detailed replication results, continued

Citation	Fixed effect	Joint test of interactions		Test of treat. diff.		
		Wald stat.	DF	p -value	t stat.	p -value
Lochner and Moretti (2004)	Race (all)	24.22	1	0.000	-4.92	0.000
	Age (blacks)	865.10	13	0.000	12.93	0.000
	Year (blacks)	41.60	2	0.000	5.69	0.000
	Age (whites)	1860.06	13	0.000	14.22	0.000
Meghir and Palme (2005)	Year (whites)	12.03	2	0.002	1.07	0.286
	High father's ed.	46.73	1	0.000	1.16	0.244
	Sex	0.40	1	0.527	0.32	0.747
	Year	41.96	11	0.000	2.49	0.013
Oreopoulos (2006)	N.Ireland	44.65	1	0.000	3.89	0.000
	Age (GB)	879.85	25	0.000	0.92	0.360
	Age (NI)	148468.65	25	0.000	-1.44	0.150
	Age (NI & GB)	173.47	28	0.000	0.48	0.634
Pérez-González (2006)	High fam. own. (MB)	1.48	1	0.223	-1.17	0.243
	High fam. own. (OR)	0.49	1	0.483	-0.69	0.489
	Year (MB)	39.78	18	0.002	0.98	0.329
	Year (OR)	34.88	18	0.010	0.22	0.829
Karlan and Zinman (2008)	Mailer wave	2.21	2	0.330	0.21	0.837
	Risk category	8.26	2	0.016	-2.57	0.010

Notes: Column 1 gives the paper and column 2 gives the fixed effects under consideration. Column 3 gives the Wald statistic of a joint test of the significance of the interactions, column 4 gives the degrees of freedom for that test, and column 5 gives the p -value. Column 6 gives a t statistic from a test of the difference between the FE and sample-weighted estimates using the test derived in Appendix 3.A.2 and the corresponding p -value.

Table 3.11: Detailed replication results for Banerjee and Iyer (2005)

Outcome	Fixed effect	FE est.	FE SE	SWE	SWE SE	% Diff.
Prop. Fertilized	Soil — red	10.71	3.33	12.03	3.47	12.4
	Soil — black	10.71	3.33	10.78	3.46	0.7
	Soil — all	10.71	3.33	10.67	3.36	-0.4
	Coastal	10.71	3.33	10.73	3.33	0.2
	Year	10.71	3.33	10.76	3.34	0.5
Log yield	Soil — red	0.16	0.07	0.17	0.07	10.3
	Soil — black	0.16	0.07	0.16	0.07	2.1
	Soil — all	0.16	0.07	0.17	0.07	5.3
	Coastal	0.16	0.07	0.16	0.07	-0.8
	Year	0.16	0.07	0.16	0.07	0.0
Log rice yield	Soil — red	0.17	0.08	0.16	0.08	-3.6
	Soil — black	0.17	0.08	0.18	0.08	4.2
	Soil — all	0.17	0.08	0.18	0.08	5.8
	Coastal	0.17	0.08	0.17	0.08	-0.5
	Year	0.17	0.08	0.17	0.08	-0.2
Log wheat yield	Soil — red	0.23	0.07	0.24	0.07	6.4
	Soil — black	0.23	0.07	0.24	0.07	3.4
	Soil — all	0.23	0.07	0.24	0.07	6.8
	Coastal	0.23	0.07	0.21	0.07	-6.7
	Year	0.23	0.07	0.23	0.07	-0.1

Notes: Column 1 gives the paper and column 2 gives the fixed effects under consideration. Columns 3 and 4 give the standard FE model estimate of the treatment effect and its standard error. Columns 5 and 6 give the sample-weighted estimate from an interacted model and its standard error. The final column gives the percent difference between the FE and SWE estimates.

Table 3.12: Detailed replication results for Banerjee and Iyer (2005), continued

Outcome	Fixed effect	FE est.	FE SE	SWE	SWE SE	% Diff.
Prop. Cereals	Soil — red	0.06	0.03	0.05	0.03	-17.1
	Soil — black	0.06	0.03	0.06	0.03	-0.2
	Soil — all	0.06	0.03	0.06	0.03	6.6
	Coastal	0.06	0.03	0.06	0.03	0.5
Prop. HYV rice	Year	0.06	0.03	0.06	0.03	0.1
	Soil — red	0.08	0.04	0.08	0.05	0.9
	Soil — black	0.08	0.04	0.08	0.04	1.1
	Soil — all	0.08	0.04	0.08	0.04	3.0
Prop. HYV wheat	Coastal	0.08	0.04	0.08	0.04	0.2
	Year	0.08	0.04	0.08	0.04	-0.2
	Soil — red	0.09	0.05	0.07	0.05	-20.5
	Soil — black	0.09	0.05	0.07	0.05	-18.3
Prop. Irrigated	Soil — all	0.09	0.05	0.09	0.05	3.3
	Coastal	0.09	0.05	0.09	0.04	-1.5
	Year	0.09	0.05	0.09	0.05	0.6
	Soil — red	0.07	0.03	0.08	0.03	19.5
Prop. Irrigated	Soil — black	0.07	0.03	0.07	0.04	4.7
	Soil — all	0.07	0.03	0.07	0.03	2.0
	Coastal	0.07	0.03	0.06	0.03	-1.1
	Year	0.07	0.03	0.07	0.03	0.0

Notes: Column 1 gives the paper and column 2 gives the fixed effects under consideration. Columns 3 and 4 give the standard FE model estimate of the treatment effect and its standard error. Columns 5 and 6 give the sample-weighted estimate from an interacted model and its standard error. The final column gives the percent difference between the FE and SWE estimates.

Table 3.13: Detailed replication results for Banerjee and Iyer (2005), continued

Outcome	Fixed effect	Joint Test of interactions		Test of treat. diff.		
		Wald stat.	DF	<i>p</i> -value	<i>t</i> stat.	<i>p</i> -value
Prop. Fertilized	Soil — red	4.52	1	0.033	1.46	0.144
	Soil — black	0.06	1	0.810	0.24	0.814
	Soil — all	0.04	1	0.848	0.18	0.857
	Coastal	0.28	1	0.598	0.19	0.846
	Year	124.52	31	0.000	0.93	0.351
Log yield	Soil — red	2.06	1	0.152	1.19	0.233
	Soil — black	0.14	1	0.711	0.35	0.724
	Soil — all	3.48	1	0.062	0.67	0.502
	Coastal	1.16	1	0.282	0.24	0.807
	Year	274.22	31	0.000	-0.88	0.378
Log rice yield	Soil — red	0.40	1	0.528	0.60	0.548
	Soil — black	1.19	1	0.276	0.67	0.501
	Soil — all	6.29	1	0.012	0.62	0.538
	Coastal	1.31	1	0.252	0.22	0.829
	Year	171.87	31	0.000	-1.05	0.294
Log wheat yield	Soil — red	1.04	1	0.308	1.21	0.225
	Soil — black	0.47	1	0.493	0.61	0.540
	Soil — all	6.97	1	0.008	0.87	0.387
	Coastal	3.05	1	0.081	1.44	0.149
	Year	117.86	31	0.000	-0.48	0.628

Notes: Column 1 gives the paper and column 2 gives the fixed effects under consideration. Column 3 gives the Wald statistic of a joint test of the significance of the interactions, column 4 gives the degrees of freedom for that test, and column 5 gives the *p*-value. Column 6 gives a *t* statistic from a test of the difference between the FE and sample-weighted estimates using the test derived in Appendix 3.A.2 and the corresponding *p*-value.

Table 3.14: Detailed replication results for Banerjee and Iyer (2005), continued

Outcome	Fixed effect	Joint Test of interactions			Test of treat. diff.	
		Wald stat.	DF	<i>p</i> -value	<i>t</i> stat.	<i>p</i> -value
Prop. Cereals	Soil — red	3.09	1	0.079	1.18	0.237
	Soil — black	0.00	1	0.973	0.03	0.973
	Soil — all	4.97	1	0.026	0.54	0.587
	Coastal	0.05	1	0.832	0.21	0.837
Prop. HYV rice	Year	78.04	22	0.000	0.18	0.854
	Soil — red	0.01	1	0.928	0.09	0.929
	Soil — black	0.04	1	0.837	0.20	0.841
	Soil — all	1.05	1	0.305	0.55	0.583
Prop. HYV wheat	Coastal	0.12	1	0.729	0.22	0.827
	Year	108.78	22	0.000	-0.62	0.536
	Soil — red	6.31	1	0.012	1.50	0.133
	Soil — black	8.02	1	0.005	1.21	0.225
Prop. Irrigated	Soil — all	2.64	1	0.104	0.46	0.649
	Coastal	7.58	1	0.006	0.16	0.873
	Year	179.01	22	0.000	0.48	0.628
	Soil — red	3.07	1	0.080	1.36	0.172
Prop. Irrigated	Soil — black	0.75	1	0.387	0.70	0.482
	Soil — all	0.35	1	0.555	0.45	0.649
	Coastal	1.43	1	0.231	0.22	0.827
	Year	84.84	26	0.000	-0.12	0.901

Notes: Column 1 gives the paper and column 2 gives the fixed effects under consideration. Column 3 gives the Wald statistic of a joint test of the significance of the interactions, column 4 gives the degrees of freedom for that test, and column 5 gives the *p*-value. Column 6 gives a *t* statistic from a test of the difference between the FE and sample-weighted estimates using the test derived in Appendix 3.A.2 and the corresponding *p*-value.

Bibliography

- Acemoglu, Daron and Joshua Angrist. 2000. “How Large are Human Capital Externalities? Evidence from Compulsory Schooling Laws.” *NBER Macroeconomics Annual* 15:9–59.
- Angrist, Joshua D. and Alan B. Krueger. 1999. Empirical Strategies in Labor Economics. In *Handbook of Labor Economics*, ed. Orley Ashenfelter and David Card. Vol. 3 Elsevier.
- Angrist, Joshua and Jörn-Steffen Pischke. 2009. *Mostly Harmless Econometrics*. Princeton University Press.
- Ariely, Dan, George Loewenstein and Drazen Prelec. 2003. ““Coherent Arbitrariness”: Stable Demand Curves Without Stable Preferences.” *The Quarterly Journal of Economics* 118(1):73–106.
- Banerjee, Abhijit and Lakshmi Iyer. 2005. “History, Institutions, and Economic Performance: The Legacy of Colonial Land Tenure Systems in India.” *American Economic Review* 95(4):1190–1213.
- Bateman, Ian J., Richard T. Carson, Michael Hanemann Brett Day and Nick Hanley. 2002. *Economic Valuation with Stated Preference Techniques: A Manual*. Elgar.
- Becker, Gordon M., Morris H. DeGroot and Jacob Marschak. 1964. “Measuring utility by a single-response sequential method.” *Behavioral Science* 9:226–236.
- Bedard, Kelly and Olivier Deschênes. 2006. “The Long-Term Impact of Military Service on Health: Evidence from World War II and Korean War Veterans.” *American Economic Review* 96(1):176–194.
- Bohm, Peter, Johan Lindén and Joakim Sonnegård. 1997. “Eliciting Reservation Prices: Becker-DeGroot-Marschak Mechanisms vs. Markets.” *The Economic Journal* 107(443):1079–1089.

BIBLIOGRAPHY

- Cameron, A. Colin and Pravin K. Trivedi. 2005. *Microeconometrics*. Cambridge University Press.
- Card, David, Carlos Dobkin and Nicole Maestas. 2008. “The Impact of Nearly Universal Insurance Coverage on Health Care Utilization: Evidence from Medicare.” *American Economic Review* 98(5):2242–2258.
- Caspar, Max. 1959. *Kepler*. Dover Books on Astronomy Dover Publications. Republished, 1993.
- Chow, Gregory C. 1960. “Tests of Equality Between Sets of Coefficients in Two Linear Regressions.” *Econometrica* 28(2):591–605.
- Clerke, Agnes Mary. 1885. *A popular history of astronomy during the nineteenth century*. A. and C. Black. Fourth edition, 1902.
- Friedberg, Leora. 1998. “Did Unilateral Divorce Raise Divorce Rates? Evidence from Panel Data.” *American Economic Review* 88(3):608–627.
- Glaeser, Edward L., David I Laibson, José A. Scheinkman and Christine L Soutter. 2000. “Measuring Trust.” *Quarterly Journal of Economics* 115(3):811–846.
- Goldberger, Arthur S. 1991. *A Course in Econometrics*. Harvard University Press.
- Graham, Bryan and Jim Powell. 2010. “Identification and Estimation of Average Partial Effects in ‘Irregular’ Correlated Random Coefficient Panel Data Models.” NBER working paper.
- Grether, David M. and Charles E. Plott. 1979. “Economic Theory of Choice and the Preference Reversal Phenomenon.” *American Economic Review* 69(4):629–638.
- Griffith, Rachel, Rupert Harrison and John Van Reenen. 2006. “How Special Is the Special Relationship? Using the Impact of U.S. R&D Spillovers on U.K. Firms as a Test of Technology Sourcing.” *American Economic Review* 96(5):1859–1875.
- Harrison, Glenn W. 1992. “Theory and Misbehavior of First-Price Auctions: Reply.” *American Economic Review* 82(5):1426–1443.
- Heath, T.L. 1913. *Aristarchus of Samos, the ancient Copernicus: a history of Greek astronomy to Aristarchus, together with Aristarchus’s Treatise on the sizes and distances of the sun and moon*. Clarendon press.

BIBLIOGRAPHY

- Heckman, James J. and V. Joseph Hotz. 1989. "Choosing Among Alternative Non-experimental Methods for Estimating the Impact of Social Programs: The Case of Manpower Training." *Journal of the American Statistical Association* 84(408):862–874.
- Heisenberg, Werner. 1927. "Über den anschaulichen Inhalt der quantentheoretischen Kinematik und Mechanik." *Zeitschrift für Physik A Hadrons and Nuclei* 43:172–198.
- Hoffman, Elizabeth, Dale J. Menkhous, Dipankar Chakravarti, Ray A. Field and Glen D. Whipple. 1993. "Using Laboratory Experimental Auctions in Marketing Research: A Case Study of New Packaging for Fresh Beef." *Marketing Science* 12:318–338.
- Holt, Charles A. and Susan K. Laury. 2002. "Risk Aversion and Incentive Effects." *American Economic Review* 92(5):1644–1655.
- Horowitz, John K. 2006. "The Becker-DeGroot-Marschak mechanism is not necessarily incentive compatible, even for non-random goods." *Economics Letters* 93:6–11.
- Kaas, Klaus Peter and Hiedrun Ruprecht. 2006. "Are the Vickrey Auction and the BDM Mechanism Really Incentive Compatible? – Empirical Results and Optimal Bidding Strategies in the Cases of Uncertain Willingness-to-pay." *Schmalenbachs Business Review* 93:37–55.
- Kahneman, Daniel, Jack L. Knetsch and Richard Thaler. 1990. "Experimental Tests of the Endowment Effect and the Coase Theorem." *Journal of Political Economy* 98(6):1325–1348.
- Karlan, Dean S. and Jonathan Zinman. 2008. "Credit Elasticities in Less-Developed Economies: Implications for Microfinance." *American Economic Review* 98(3):1040–1068.
- Karni, E. and Z. Safra. 1987. "Preference reversals and the observability of preferences by experimental methods." *Econometrica* 55:675–685.
- Keller, L. Robin, Uzi Segal and Tan Wang. 1993. "The Becker-DeGroot-Marschak Mechanism and Generalized Utility Theories: Theoretical Predictions and Empirical Observations." *Theory and Decision* 34:83–97.
- Kőszegi, Botond and Matthew Rabin. 2006. "A Model of Reference-Dependent Preferences." *Quarterly Journal of Economics* 121(4):1133–1165.

BIBLIOGRAPHY

- Kőszegi, Botond and Matthew Rabin. 2007. "Reference-Dependent Risk Attitudes." *American Economic Review* 97(4):1047–1073.
- Lichtenstein, Sarah and Paul Slovic. 1971. "Reversals of Preferences Between Bids and Choices in Gambling Decisions." *Journal of Experimental Psychology* 89:46–55.
- List, John A. 2003. "Does Market Experience Eliminate Market Anomalies?" *Quarterly Journal of Economics* 118(1):41–71.
- List, John A. and C. A. Gallet. 2001. "What Experimental Protocol Influence Disparities Between Actual and Hypothetical Stated Values? Evidence from a Meta-Analysis." *Environmental and Resource Economics* 20(3):241–254.
- Lochner, Lance and Enrico Moretti. 2004. "The Effect of Education on Crime: Evidence from Prison Inmates, Arrests, and Self-Reports." *American Economic Review* 94(1):155–189.
- Lochner, Lance and Enrico Moretti. 2011. "Estimating and Testing Non-Linear Models Using Instrumental Variables." NBER working paper.
- Lusk, Jayson L., Corinne Alexander and Matthew C. Rousu. 2007. "Designing Experimental Auctions for Marketing Research: The Effect of Values, Distributions, and Mechanisms on Incentives for Truthful Bidding." *Review of Marketing Science* 5(3).
- Machina, Mark J. 1987. "Choice Under Uncertainty: Problems Solved and Unsolved." *The Journal of Economic Perspectives* 1(1):121–154.
- Mazar, Nina, Botond Kőszegi and Dan Ariely. 2009. "Price-Sensitive Preferences." Mimeo.
- Meghir, Costas and Marten Palme. 2005. "Educational Reform, Ability, and Family Background." *American Economic Review* 95(1):414–424.
- Noussair, Charles, Stephane Robin and Bernard Ruffieux. 2004. "Revealing consumers' willingness-to-pay: a comparison of the BDM mechanism and the Vickrey auction." *Journal of Economic Psychology* 25(6):725–741.
- Oreopoulos, Philip. 2006. "Estimating Average and Local Average Treatment Effects of Education when Compulsory Schooling Laws Really Matter." *American Economic Review* 96(1):152–175.
- Papke, Leslie E. 1994. "Tax Policy and Urban Development: Evidence from the Indiana Enterprise Zone Program." *Journal of Public Economics* 54:37–49.

BIBLIOGRAPHY

- Pérez-González, Francisco. 2006. "Inherited Control and Firm Performance." *American Economic Review* 96(5):1559–1588.
- Plott, Charles and Kathryn Zeiler. 2005. "The Willingness-to-Pay–Willingness to Accept Gap, the Endowment Effect, Subject Misconceptions, and Experimental Procedures for Eliciting Valuations Effect." *American Economic Review* 95(3):530–545.
- Rawlins, Dennis. 1993. "Tycho's 1004 Star Catalog." *DIO: The International Journal of Scientific History* 3(70):1–106.
- Safra, Zvi, Uzi Segal and Avia Spivak. 1990. "The Becker-DeGroot-Marschak Mechanism and Unexpected Utility: A Testable Approach." *Journal of Risk and Uncertainty* 3:117–190.
- Tversky, Amos and Daniel Kahneman. 1974. "Judgment under Uncertainty: Heuristics and Biases." *Science* 185(4157):1124–1131.
- Vickrey, William. 1961. "Counterspeculation, Auctions, and Competitive Sealed Tenders." *Journal of Economic Psychology* 16:8–37.
- Wertebroch, Klaus and Bernd Skiera. 2002. "Measuring Consumers' Willingness to Pay at the Point of Purchase." *Journal of Marketing Research* 39:228–241.
- Wooldridge, Jeffrey M. 2005a. "Fixed-Effects and Related Estimators for Correlated Random-Coefficient and Treatment-Effect Panel Data Models." *Review of Economics and Statistics* 87(2):385–390.
- Wooldridge, Jeffrey M. 2005b. Unobserved Heterogeneity and Estimation of Average Partial Effects. In *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*, ed. Donald Andrews and James Stock. Cambridge University Press.