

The Measurement, Estimation and Analysis of Subjective Probability Distributions

With Applications to Production and
Investment Decisions in Rural Tanzania

A Dissertation

Presented to the Faculty of the Graduate School
of Cornell University

in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy

by

Brian M. Dillon

August 22, 2011

©2011 Brian Dillon

Abstract

The research presented in this dissertation focuses on the measurement and analysis of subjective probability distributions over stochastic outcomes, a central issue in the study of decision-making under uncertainty. The empirical setting is rural Tanzania, where the degrees of risk and uncertainty characterizing both human capital and productive investment decisions are exacerbated by widespread dependence on rain-fed agriculture, inadequate social safety nets, and a poorly developed information infrastructure. I present a sequence of methodological, theoretical and empirical chapters in which I estimate subjective returns distributions in an existing data set, develop and explain a new method of collecting subjective distributions data, characterize the information content of the data collected, and make use of the data to estimate a structural agricultural production model.

Chapter 1 explores the role of estimated, rather than measured, subjective returns to education in schooling choice decisions. Using an existing panel survey from Tanzania, I estimate earnings-education distributions separately for 1991, 2004 and 2010. I then use individual-level predictions of the first two moments of the earnings distribution to estimate a random effects probit model on binary enrollment decisions for school-aged children in the years 1991-1994. I find that the returns to education have been and remain high for women, while for men the returns increased over the twenty study years to nearly match those of women. In addition, the probability of enrollment is increasing in the subjective conditional expectation of earnings, and decreasing in the subjective conditional variance of earnings.

Chapter 2 describes the phone-based survey method that I used to gather

subjective probability distributions data from a sample of Tanzanian cotton farmers. I describe the various technical issues faced in the implementation of this method, outline the lessons learned and the numerous refinements made over the course of the study, and speculate on the feasibility of phone-based data collection in other settings in low income countries.

In Chapter 3, I analyze the information content of subjective distributions data gathered in the way that has become standard in development economics, i.e., by having respondents allocate a fixed number of counters to boxes that represent the intervals of a histogram. I use inference about the respondents' choice problem to analyze the partial identification of the underlying belief. I provide bounds on the density in subsets of intervals, provide bounds on the underlying CDF, define the joint identification region for the measure vector, and develop and implement a feasible numerical method for jointly bounding the moments of the unobserved distribution. I also provide simulation evidence for the optimal design of survey instruments and the optimal way to approximate these data with smooth distributions.

Lastly, Chapter 4 makes use of the regularly spaced within-season measures of subjective yield and price distributions collected from Tanzanian cotton farmers to study the farmer's dynamic resource allocation problem. Using these data, I develop a novel method for estimating a stochastic production function when error parameters are observed at the plot-level throughout the cultivation season.

BIOGRAPHICAL SKETCH

Brian Dillon received an M.Phil in economics from the University of Cambridge in 2005, a joint BA in philosophy and BS in mathematics from Loyola University Chicago in 2000, and a high school diploma from Lansing Catholic Central High School in 1996.

ACKNOWLEDGMENTS

Of the many people who have played an important role in my formation as a researcher and economist, my parents Patrick and Maureen Dillon are first and foremost. They raised my sisters and me in a loving household with a high value on education, and they provided a model of hard work and intellectual curiosity that has shaped my view of the world. Through my many years of travel, learning and graduate training, they have been unwavering in their support. In economic parlance, it is fair to say that the unobservable contribution to my educational outcomes has been, in many ways, more substantial than the observable. Mom, that's an economist's way of saying, "Thanks, I love you guys."

I would also like to acknowledge the best piece of educational advice I received, which was from my father. In the summer before I began studying at Loyola, he suggested that I consider a second major in mathematics. My parents supported my decision to study philosophy, but my father noted that math often proved useful in a wide range of settings, and that since I enjoyed it anyway, I may as well double major. How right he was - there is no chance that I would be where I am today if I had not heeded this piece of advice.

My best decision in graduate school was to ask Chris Barrett to be my advisor, and my biggest stroke of fortune was that he accepted. From my first year onwards, Chris has provided in a gradually evolving sequence: formal training in the classroom, an orientation toward the research frontier, professional advice, feedback on my work, assistantship funding, the opportunity for joint work, and finally post-doctoral funding and the prospect of future collaboration. Chris has created an atmosphere of cooperation, communication and support among his graduate students and post-docs that sets him apart from his colleagues. I cannot thank him enough for both the technical skills and the broad perspective on research that I have gained under his tutelage.

I am also very grateful for the invaluable professional and personal support I received from Ted O'Donoghue and Francesca Molinari, my committee members. Through the courses they offered, the countless one-on-one meetings in which they entertained my half-formed research ideas and nudged me back in the right direction, and the examples they offered as top notch researchers, Ted and Francesca provided me with both the tools of the trade and a model of how best to use them. They have held my work to a high standard, and have challenged me to do the same.

Many other economists played important roles in my graduate training, for which I am very thankful. Kaushik Basu has been a friend, a teacher and a mentor. John Abowd and Ravi Kanbur provided critical advice, feedback on my work, and letters of recommendation. Joachim de Weerd and Mujo Moyo hosted me in Tanzania and provided invaluable guidance for my field work, not to mention great conversations around the dinner table. Kathleen Beegle opened up numerous doors for me. Kathleen hired me for my first research job in Tanzania, and, with her colleague Gero Carletto, provided collaborative funding through the LSMS-ISA division in the World Bank Development Research Group that helped make my field work possible.

I am grateful for the support of numerous other funding bodies. The research presented here was made possible by grants from the Mario Einadui Center for International Studies, the Cornell College of Arts and Sciences, the Cornell Institute for Social Sciences through its Persistent Poverty and Upward Mobility special themed project, the National Science Foundation through a Doctoral Dissertation Improvement Grant (SES-0921833), and a Chester O. McCorkle Fellowship from the Agricultural and Applied Economics Association. Responsibility for the results of the research, however, rests solely with me. The Knight Institute for Writing in the Disciplines, and the Math Department through the efforts of Maria Terrell, provided invaluable support through teaching assistantships.

In Tanzania, the staff of Economic Development Initiatives, Ltd., in Bukoba, were wonderful friends, hosts and colleagues. I am especially grateful to Aris Mgothamwende, Respichius Mitti, Benjamin Kamukulu, Thaddeus Rweyemamu, Bahati Julius, and Ma Anna. The group of enumerators, drivers and data entry technicians who helped put together the REAP data set includes Michael Kamukulu, Lilian Rugambwa, Joanitha Balthazari, Amwesiga Bandio, Michael Filbert, Anko, Zanibibi Yazid, Allen Bantanuka, Warda Maulid, and Justina. Thanks for all of your hard work. Lastly, but most importantly, I am grateful for the tireless work of Diego Shirima, Geoffrey Mwemezi and Msafiri Msedi. They played essential roles in the face-to-face and phone survey portions of REAP, and it is fair to say that I could not have put together the data for this dissertation without them.

In addition to those already listed, many other people provided helpful feedback on the work in this dissertation, whether at conferences and Cornell seminars, in conversation, or by reading drafts. These include in no particular order: Annemie Maertens, Russell Toth, Kira Marie Villa, Marc Rockmore, Aurelie Harou, Joanna Upton, Hans Hoogeveen, Alessandro Romeo,

Diane Steele, Mark Davies, Travis Lybbert, Jörg Stoye, Larry Blume, Nick Kiefer, George Jakubson, Yongmiao Hong, Alex Rees-Jones, Amalavoyal Chari, Chayanee Chawanote, Hope Michelson, Corey Lang, Tom Walker, Nishith Prakash, Kalle Hirvonen, participants at the 2010 BREAD/AMID summer school at LSE, and participants at the 2011 Econometric Society Summer Meetings.

Countless Cornell staff members have helped me in various ways, including Amy Moesch, Sheri van Deusen, Ulrike Kroeller, Erin Lentz, Gail Keenan, Rose Hastings, Linda Harris, Diane Edwards, and Jo Schroeder. Thanks to all of you.

I have been lucky to have wonderful teachers and mentors at every stage of my education. These people helped form the foundation upon which my final graduate education was built, and I am grateful to all of them: Rich Ball, Kathy and Fred Partlow, Jan Shoemaker, Fred Diehl, Becky Corner, Tom Carson, Ardis Collins, Alan Saleski, Tom Prendergast, Ken Hori, Vasilis Sarafidis, Donald Robertson, Willy Brown, and Jonathan Pincus.

Among my friends and peers at Cornell, Peter Brummund and Kurt Lavetti stand out as friends, cycling partners, and colleagues with a shared enthusiasm for research. We have grown together over our years at Cornell, and I look forward to decades of friendship and professional collaboration. Of the many other wonderful friends who have provided support of various kinds over the past five years, I am especially grateful to have in my life Robin Kachka, Annelies Deuss, Sergio Pulido, Michelle McCutcheon-Schour, Reggie Covington, Karen Brummund, Jim Casteleiro, Sonali Das, Elena Solodow, and Eric Maroney.

Finally, I am grateful to, Anya, who provided constant love and support through some difficult and stressful years of graduate school. I'm very lucky to have you in my life, and I couldn't have done this without you. Along with Eloise and Zoe, of course.

Contents

1	Introduction	1
2	Subjective Returns to Education	11
2.1	Introduction	11
2.2	Data	17
2.3	Estimating the Effect of Education on Earnings	21
2.3.1	Results: earnings and education	26
2.3.2	Estimating the variance of log earnings	30
2.3.3	Predicting the expectation and variance of earnings in levels	34
2.4	Enrollment Decisions	41
2.4.1	Results of random effects probit model	43
2.5	Conclusion	47
2.6	Appendix	49
3	Using Phones to Collect Data	67
3.1	Introduction	67
3.2	REAP Project Description	68
3.3	Challenges, Solutions and Lessons Learned	71
3.3.1	Costs	72
3.3.2	Infrastructure Issues	73
3.3.3	Selection and Participation	75
3.3.4	Data Quality	80
3.3.5	Replacement of Materials	83
3.4	Conclusion	85
4	Identification of Underlying Beliefs	87
4.1	Introduction	87

4.2	The Respondent's Choice Problem	94
4.2.1	Preliminaries	94
4.2.2	Four Allocation Heuristics	97
4.2.3	Implications of Respondents' Decision Rule	102
4.3	Bounds on Bin Probabilities	105
4.3.1	Single bin bounds	106
4.3.2	Bounds on subsets of bins	116
4.3.3	Bounds on the median of f	122
4.3.4	Joint Identification Region for p	122
4.4	Bounds on the Expectation and Variance	128
4.4.1	Bounds on the expectation of f	129
4.4.2	Joint bounds on the expectation and variance	132
4.5	Simulation results	140
4.5.1	<i>Ex ante</i> choice of N and k	140
4.5.2	<i>Ex post</i> smoothing	142
4.6	Conclusion	148
4.7	Appendix	149
5	Dynamic Production Model	162
5.1	Introduction	162
5.2	Data and Setting	166
5.3	Stochastic Production Model	172
5.3.1	Timing	175
5.4	Estimation Procedure	177
5.4.1	Identification of error density functions	177
5.4.2	Estimation Algorithm	186
5.4.3	Standard errors	188
5.5	Results	189
5.6	Conclusion	191
5.7	Appendix	193
6	References	207

Chapter 1

Introduction

Economists generally believe that appropriately regulated markets, operating with the support of sufficient physical and institutional infrastructure, do a better job of allocating capital and channeling goods and services to households than do large-scale government planning programs. In recent decades, this belief has underpinned a historically unparalleled shift away from government intervention and toward market liberalization in most of the developing world. In retrospect, that shift relied more on theoretical assumptions than on hard empirical evidence about how markets function, especially in rural areas of low-income countries. In this dissertation I address this broader issue by focusing on a foundational microeconomic topic that is critical to market operation, both in the empirical setting of rural East Africa and elsewhere: agents' formation of and response to subjective probability distributions over uncertain future outcomes.

It has long been known that subjective expectations matter as much as

preferences in determining the choices people make.¹ Subjective beliefs play a particularly important role in investment and production decisions in rural East Africa, where uncertainty has an outsized effect on household welfare. Consumption in this environment is closely linked to the outputs of subsistence and cash crop agriculture, rain-fed production processes that are highly stochastic and subject to a long lag between investment and outcome. Information systems are inadequate or lacking altogether, preventing widespread knowledge dissemination and efficient updating of priors. And, lastly, the costs associated with misinformation or incorrect beliefs are relatively higher than in wealthy countries: mis-allocation of productive resources due to incorrect subjective expectations can lead to catastrophic welfare losses for households struggling to meet subsistence requirements.

Despite the prominent role of subjective probability distributions in resource allocation and investment decisions, the collection and analysis of subjective distributional data was professionally taboo for most of the 20th century. It has only recently become acceptable for economists to gather expectations data and treat them with the same sanctity as choice data. The reasons for this bias against expectations data are not entirely clear, though they are most likely related to the rise of rational expectations theory in the 1970s, and the perception that choice data is in some sense more objective than reported expectations. Manski (2004) speculates on these issues at some length.

¹For a classic example, see Marshall (1895).

Regardless of the true history, there are exciting new strains of research that attempt to measure agents' expectations, understand how they are formed, and investigate their importance for decision making. In Malawi, Delavande and Kohler (2008a, 2008b) study expectations over events such as HIV contraction, market participation and experience of shocks using both a simple Likert scale (Very likely, Somewhat likely, Likely, etc.) for point estimation of expectations, and more sophisticated methods of eliciting distributions by dividing a fixed number of counters into piles representing various outcomes. The method of dividing counters such as stones or beans into piles representing the likelihood of different outcomes is also used by Hill (2010) in Uganda to collect expectations about coffee prices, by Lybbert et al (2007) and Luseno et al (2003) to collect Kenyan and Ethiopian pastoralists' expectations about rainfall in the coming season, and by Santos and Barrett (2010) to elicit state-conditional herd size distributions in southern Ethiopia. Cole and Hunt (2010) and Camacho and Conover (2011) study price expectations, and find in separate studies that provision of price data to farmers via SMS significantly improves the accuracy of subjective price distributions, but has no measurable effect on production practices.

The work presented here makes important theoretical and empirical contributions to this literature. The data underlying all but the first empirical chapter of this dissertation came from a 16-month study of cotton farmers in Tanzania. The Tanzanian cotton sector is something of a flagship for market liberalization in Africa. The sector exhibits a number of characteristics

that make it especially suitable for a study of investment under uncertainty. Cotton is an inedible cash crop, so that producer objectives are not complicated by the possibility of consuming their output. Cotton takes six months to grow, and growth over the cultivation period depends critically on both chosen inputs and on the outcomes of stochastic rainfall and pest processes. Farmers' expectations about these processes figure substantially in their resource allocation decisions. Also, cotton prices are entirely unsecured and are unknown to farmers for the majority of the growing season. The value of output is therefore a random variable over which there is significant heterogeneity in beliefs.

The last two decades have seen the near complete retreat of the Tanzanian government from the sector, reducing rent-seeking and some forms of inefficiency, while simultaneously increasing the exposure of farmers to the vagaries of the world market. The collapse of the government-sponsored input allocation system has left cotton growers more susceptible to drought, severe weather events and pest infestation. In this atmosphere of considerable uncertainty, nearly half a million small scale Tanzanian farmers cultivate cotton without the support of formal insurance institutions, sophisticated information services or publicly funded social safety nets, all of which would serve to mitigate the negative effects of uncertainty. The choice behavior of these farmers, who produce for what is arguably the most liberalized output market in East Africa, offers an important window into the choice behavior of all producers who operate at the intersection of centuries old subsistence

traditions and modern, globalized markets.

The data set collected for this project focused on subjective distributions. From July 2009 until November 2010, we gathered both high frequency input allocation data and subjective probability distributions over all of the major sources of uncertainty in the crop revenue function: prices, yields, rainfall and pest pressure. From the pre-planting period until the final sale of cotton, we observe both choice data and individual-level subjective beliefs. We use these data to better understand the role of uncertainty in agricultural production in Tanzania, and to improve estimation of structural agricultural production functions.

The emphasis on agriculture connects this dissertation to the broad literature on the fundamental role of expectations and risk attitudes in agricultural activity. Agricultural production is inherently forward-thinking, as investments are usually undertaken many months or even years in advance of outcomes, with little scope for recovery or reversibility of sunk costs. Output is subject to a high degree of uncertainty, stemming from price changes and the effect of exogenous and/or unobserved biological inputs. Moschini and Hennessy (2001) survey these issues in detail, and Chavas (2004) covers the connection between theoretical models of risk and empirical specifications for testing predictions. Because subsistence agriculture and home production play such a fundamental role in the lives of people in low income, agrarian countries, and because realizations of the downside risks facing poor households often lead to dire consequences for health and the probability of

survival, the development economics literature has long been concerned with the role of risk and expectations in individual behavior. Nerlove and Bessler (2001) review the primary methods used to recover expectations from agricultural choice data, including structural models of adaptive, implicit and rational expectations, and empirical time series models. Absent from their work is any discussion of directly measured subjective expectations, precisely because this line of work is new.

With choice under uncertainty and the analysis of subjective probability distributions as the thematic backdrop, this dissertation proceeds in four chapters. In Chapter 2, “Subjective Returns to Education and Schooling Choice: Evidence from Tanzania”, we use an existing panel survey from northwest Tanzania to estimate subjective distributions of the returns to education for children for whom parents make school enrollment decisions in the early 1990s. We estimate subjective returns by looking at the future, realized outcomes of adults similar to the school-aged children from the early 1990s. The first two moments of the estimated subjective returns distributions are then used to study the original enrollment decision. While the findings are in line with our priors and are robust up to the chapter’s numerous parametric assumptions, this chapter demonstrates the identification problems and heavy parametric burden placed on even a detailed micro data set when subjective distributions are estimated from choice data rather than observed directly.

Chapter 3, titled “Using Mobile Phones to Collect Panel Data in De-

veloping Countries”, which is forthcoming in the *Journal of International Development*, describes the methods used to gather subjective probability distributions data from the sample of Tanzanian cotton farmers. High frequency data are required in order to study the role of subjective beliefs in a stochastic control process that persists for over six months and is subject to the gradual revelation of information about states of nature. Lacking resources to embed enumerators in numerous remote villages for months at a time, we opted instead to gather high frequency survey data by mobile phone. This novel data collection method only recently became possible, with the rapid proliferation of mobile network coverage throughout rural northwest Tanzania. Chapter 2 describes the various technical issues faced in the implementation of this method. We outline the lessons learned and the numerous refinements made over the course of the study, and speculate on the feasibility of phone-based data collection in other settings in low income countries. For dissertation purposes, Chapter 2 provides a record of the field work undertaken.

In Chapter 4, “Identification of Underlying Beliefs from Subjective Distributions Data”, we analyze the information content of subjective distributions data gathered in the fashion that has quickly become standard in development economics, i.e., by having respondents allocate a fixed number of counters to boxes on a visual aid that represent the intervals of a histogram. We first argue that regardless of the method used to gather subjective distributions data, all such data should be viewed as the outcome of a choice

problem solved by respondents. This is because it is impossible for even fully informed and honest respondents to communicate their infinite-dimensional probabilistic beliefs to researchers. In this chapter we use inference about the respondents' choice problem to analyze the partial identification of the underlying belief. We provide bounds on the density in any subset of the intervals, provide bounds on the underlying CDF, define the joint identification region for the measure vector, and describe a method for jointly bounding the moments of the unobserved distribution. We also provide simulation evidence for the optimal design of survey instruments and the optimal way to approximate these data with smooth distributions.

Lastly, Chapter 5 makes use of the regularly spaced within-season measures of subjective yield and price distributions collected from Tanzanian cotton farmers to study the farmer's dynamic resource allocation problem. Using these data, we develop a novel method for estimating a stochastic production function when error parameters are observed at the plot-level throughout the cultivation season. We then use the estimated model parameters to calculate the value of price uncertainty, defined as the profit loss from having incorrect beliefs about the end-of-season output price distribution, for every farmer, at each major stage of cultivation. Results are compared with results based on the method in Fafchamps (1993), which studied a similar problem without the benefit of subjective distributions data.

In broad terms, the work presented here brings together three lines of inquiry. We contribute to the literature on market liberalization by directly

confronting the issues of information transmission and expectations formation in a rural, low income setting. We study farmer choice under uncertainty, an issue with a long empirical and theoretical tradition in agricultural economics, using a novel data set in which subjective distributions are observed in a high frequency panel, rather than modeled and estimated by the researcher. Lastly, we add to the rapidly growing literature on the measurement and analysis of subjective expectations by studying the evolution of farmer's beliefs over a full cultivation cycle, and by characterizing the partial identification problem that every researcher faces when dealing with data of this kind.

If we take standard microeconomic theory seriously, we know that choice data in situations characterized by uncertainty is best understood if accompanied by a density function that can be indexed at the individual and time period level with actual data, rather than with distributions generated by the researcher using highly restrictive, untestable assumptions. This dissertation, then, represents an attempt to take seriously the issues of heterogeneity and inter-temporal variation in subjective probability distributions, with specific applications in rural East Africa. We provide empirical results from an existing data set in which expectations are not observed, methodological findings that pave the way for high frequency collection of distributional data, econometric theory that characterizes the information content of the data collected in this and numerous other studies, and structural estimates of the parameters of a dynamic agricultural production model that make use

of the observed inter-personal and inter-temporal variation in beliefs. To our knowledge, the work described in this dissertation is the first attempt to go beyond static measures of subjective expectations, and investigate the evolution of full subjective distributions over time. As such it represents an important contribution to the new literature on subjective expectations, and to the longstanding issue of agricultural production under uncertainty.

Chapter 2

Conditional Returns to Education and Schooling Choice: Evidence from Tanzania

2.1 Introduction

Human capital theory suggests an important role for education in both promoting growth at the macro level and improving earnings and welfare outcomes at the individual level. An extensive body of empirical development work has studied the effect of educational attainment on cognitive skills, wages, earnings from self-employment, and other outcomes such as fertility

and health (Glewwe 2002, Glewwe and Kremer 2005). Additional work has separately studied the effect of cognitive skills, when separately identifiable from educational attainment, on earnings and other outcomes (Boissiere *et al* 1985). This literature is guided by a clear policy goal: to identify the education policies most likely to raise incomes and improve welfare in less developed countries.

A separate, but clearly related, goal, is to better understand the choice behavior of households when making educational choices for school-aged children. A small body of theoretical and empirical literature on the linked fertility and education decisions explicitly models and tests parents' educational choices for their children (Ejrnæs and Portner 2004). Given that the estimated private return to education in low income countries is typically very high, particularly the returns to completion of primary and secondary school, it is an open question as to why more households in poor countries do not undertake greater educational investments. One candidate explanation for sub-optimal investment in education is widespread misperception of the conditional returns distribution. Recent experimental work by Jensen (2010) and Nguyen (2008) demonstrates that providing information to students in the form of average returns or anecdotes increases enrollment and attendance.

There are other explanations for seemingly sub-optimal levels of educational investment, including credit constraints and high implicit costs. In this paper we explore the possibility that households may actually be making

reasonable choices given the subjective returns distributions that they face. When making an enrollment choice for a child, household members have information about the child's ability that is unobserved to the researcher. In this paper we make use of a panel of children whose parents made enrollment decisions in successive years in the early 1990s to control for these unobserved effects. We explicitly model and estimate the marginal conditional return to education 10 and 16 years after the observed educational investments. We then use the predictable components of the first two moments of the conditional return distribution to study the original educational investment. In essence, we treat both contemporaneous and future conditional earnings distributions as candidate subjective distributions of educational returns, and compare each distribution's ability to explain educational investments. The paper's broad contribution is to take seriously the household's perceived child-specific conditional return to education, so as to better inform policies that seek to increase educational investments and/or the returns to education in a particular sector.

Throughout this paper we maintain the assumption that the subjective, or perceived, returns to education are equivalent to the individually heterogeneous, conditional returns to education that we estimate using observed individual characteristics. This is admittedly a strong assumption, requiring not only a well-specified and fully identified returns equation, but also observation of all of the primary determinants of the earnings-education distribution. However, it represents a substantial loosening of the restrictive

assumption of a constant return to education across persons and marginal education years, which is implicit in many existing studies of educational returns. Therefore, from here on we use the term “subjective” to refer to the objective, estimated conditional returns that we estimate in the first half of the paper.

To address the questions of interest in this paper we have to overcome two identification problems that are endemic to studies of education. First is the difficult task of unbiased estimation of the effect of education on earnings. Card (1999) discusses the problems in great detail; the primary issue is the endogeneity of education, as both earnings and education are surely determined in part by a measure of ability that is unobserved to the researchers. We deal with this by instrumenting for education using the only instrument available in our data, father’s education. In developing countries, researchers often estimate the effect of education on earnings using samples consisting of only formal sector wage-earners, despite the fact that the majority of persons in low income countries do not make use of their education in a waged job. The resulting selection bias often goes uncorrected due to limitations in the data. In this paper, by utilizing all age-eligible workers in the data regardless of sector of employment, we mitigate any selection problem at the earnings estimation level.

The second identification problem relates to the schooling choice decision.¹ Glewwe and Kremer (2005) describe in detail the barriers to unbiased

¹By which we mean the choice of the number of years of education to acquire, not the

estimation of the determinants of school choice. They posit that without complete observation of all child, community and school characteristics, as well as relevant price vectors and education policies, studies that seek to explain schooling choice are sure to suffer from omitted variable bias. They advocate for the use of natural experiments (as in Duflo, 2001) and randomized control trials (as in Glewwe *et al*, 2009) to overcome these challenges. However, essentially all of the estimation problems identified by Glewwe and Kremer (2005) can be overcome with panel data, if we accept the probability of enrollment in the marginal grade, rather than the total number of years of schooling chosen, as the outcome variable of interest. With panel data we can control for the effect of time-invariant individual, household, community and school characteristics, as well as any observed time-varying determinants of enrollment (such as costs).

Our interest in this paper is in two sets of hypotheses. The first relates to the subjective perception of household members when they make educational decisions. If the distribution of returns to education is non-stationary, then the parents' capacity to forecast changes in the distribution when making educational choices has a significant effect on future child welfare. Over the course of the 1990s, the Tanzanian education system underwent a substantial overhaul at the behest of donor nations and international organizations. Universal primary education became an explicit goal of the central government, and numerous education policies were changed. We test the extent to which choice of a particular school from a menu of options.

the earnings-experience-education distribution may have changed between 1991, the earliest year of data, and 2010, the most recent, by estimating the first two moments of the conditional returns distribution separately in 1991, 2004 and 2010. If the distributions of conditional earnings changed substantially over the twenty survey years, educational investment decisions made using the contemporaneous (i.e., the early 1990s) distribution will be sub-optimal.

Second, we explore the degree to which both the expectation and the variance of the subjective returns distribution explain school choice. While we have omitted a stylized choice model from the paper in order to focus on the numerous estimation results, at the heart of the paper is a standard investment model with utility concave in wealth. We hypothesize that the children with higher subjective variances of returns - i.e. children who look like people who grow up to have high variance future earnings, based on characteristics known at the time of educational investments - are less likely to invest in additional school. This hypothesis follows naturally from a standard assumption of risk aversion over future earnings in an educational investment model. Obviously, we also hypothesize that the higher the first moment of the subjective returns distribution, conditional on other determinants of enrollment, the greater the probability of enrollment in school.

It is our particular interest in the second moment of the returns distribution that merits the parametric approach in the first half of the paper, in which we estimate the returns in the three cross-sectional years for which

we have adequate data. If we were only interested in expected earnings, we could use average realized earnings within subgroups as an estimator of the first moment of the subjective returns distribution. But to do so would be to ignore information on the stochastic component of returns, i.e. the distributional information contained in the 1991, 2004 and 2010 cross-sections. Instead, we parameterize the variance of earnings, so as to have a means of forecasting its predictable component. If we had a very large sample, we could perhaps do this non-parametrically, by using the observed mean and variance of earnings within age-education-gender subgroups. But these cells are too sparsely populated in the data, so that the within-cell variance of earnings is surely subject to substantial sampling error.

The paper proceeds as follows. In the next section we discuss the data. In Section 2.3 we discuss possible ways of specifying Mincerian returns functions, given the data limitations and requirements for the second stage, and we present the empirical results for each of the three cross-sectional years of interest. Section 2.4 presents the empirical model and the results of the second stage of the paper, in which we study enrollment decisions. Section 2.5 concludes.

2.2 Data

Data for this project are from the Kagera Health and Development Survey (KHDS). From 1991-1994 the KHDS collected four waves of panel data from

912 households² and 6,204 individuals in northwest Tanzania. In 2004 a follow-up survey was conducted, with 4,441 of the original respondents successfully re-interviewed, as well as all members of households that they had formed or joined in the intervening 10 years. An additional round of follow-up data was gathered in 2010. The KHDS gathered standard household data, including demographics, health and anthropometry, education, time use and labor experience, assets, land, agricultural production, livestock, credit and remittances, migration, consumption, income, and shocks. Most modules were enumerated in all survey years, however some key data were not gathered in 2010. We discuss this further below.

Table 2.1 provides summary statistics for the sample of school-aged children from each wave, as well as for the subset of children who appear in all four waves. The table is weighted at the individual level, thus the characteristics of households with greater numbers of school-aged children are disproportionately represented. The sample is almost exactly 50% male. The average respondent age is a little over 13 and increases slightly over the course of the survey. Birth order is defined with respect to all household members under the age of 30, because some household heads have no direct children, and because the presence in the household of young adults who are no longer school-aged could have some bearing on the educational status of school-aged youth. Literacy of the household head is defined as the stated ability to both read and write a letter. Household size grows steadily over the

²759 households were interviewed in all four rounds.

course of the survey, though this may be due to a combination of births, entry into the household and enumerator efforts to list all previous household members even if they have left the household. Average household employment income is very low in all four survey waves, largely because many households report zero outside employment. The average number of household members working on the farm is over 4 in all years, about half of household members. Average numbers of household members working in the market or in self-employment are much lower, around 0.45 and 0.26, respectively. Average hours worked per household member per sector falls across all survey waves. This is almost certainly an artifact of the increasing household size.

Comprehensive earnings data were not gathered in all rounds of the KHDS. We therefore impute individual earnings from household consumption data, using the proportional share of income contributed by (gender)-(age)-(status in the household) subgroups in the 2004 cross-section. Methodological details are discussed in the Appendix. Given that the consumption recall period in rounds 2-4 differed from that used in 1991, 2004 and 2010, we use only the latter three years to construct earnings-education-experience distributions.

Table 2.1: Summary statistics for age-eligible individuals in 1991-1994

	Wave 1	Wave 2	Wave 3	Wave 4
Male	0.50 (0.5)	0.50 (0.5)	0.50 (0.5)	0.51 (0.5)
Age in years	13.07 (4.2)	13.25 (4.2)	13.27 (4.3)	13.42 (4.2)
Birth order among those <= 30 yrs	3.18 (1.9)	3.19 (2)	3.18 (1.9)	3.10 (1.9)
Muslim	0.13 (0.3)	0.12 (0.3)	0.12 (0.3)	0.12 (0.3)
Catholic	0.59 (0.5)	0.61 (0.5)	0.61 (0.5)	0.61 (0.5)
Christian	0.24 (0.4)	0.24 (0.4)	0.24 (0.4)	0.25 (0.4)
Number of HH businesses	0.38 (0.7)	0.63 (0.8)	0.80 (1)	0.92 (1.2)
Tropical livestock units	1.81 (7.2)	2.22 (8.1)	2.27 (8.7)	2.25 (10.2)
Age of head	49.47 (15.8)	50.56 (15.8)	50.80 (15.1)	51.09 (14.6)
Head is literate	1.30 (0.5)	1.29 (0.5)	1.28 (0.5)	1.27 (0.4)
Acres cultivated	5.44 (4.8)	5.87 (5.4)	5.85 (5)	5.87 (5)
Asset index	0.08 (0.9)	0.07 (1)	0.04 (0.9)	0.06 (0.9)
HH size	8.01 (3.6)	8.57 (4.1)	9.14 (3.9)	9.68 (4)
HH annual employment income (TZS)	26869.36 (85268.1)	11690.14 (31568.4)	12215.20 (33403.1)	13203.47 (36504.2)
Number HH members working for govt	0.13 (0.4)	0.11 (0.4)	0.10 (0.3)	0.09 (0.3)
Farm: total number of HH workers	4.26 (2.3)	4.45 (2.3)	4.84 (2.4)	4.51 (2.3)
Market: total number of HH workers	0.42 (0.7)	0.44 (0.7)	0.50 (0.8)	0.46 (0.8)
Self-emp: total number of HH workers	0.28 (0.7)	0.28 (0.7)	0.25 (0.6)	0.26 (0.6)
Ag work hours per HH member	44.47 (30.6)	35.43 (24.7)	34.36 (22.2)	29.40 (20)
Mkt work hours per HH member	7.52 (17.8)	7.28 (15.4)	6.53 (14.6)	5.88 (13)
Self work hours per HH member	3.68 (10.9)	3.06 (9.2)	2.66 (8.3)	2.45 (7.8)
Agricultural equipment (TZS)	1369.82 (5167.2)	21799.67 (190371.2)	9180.65 (27851.6)	7505.90 (21685.5)
N	1388	1405	1407	1305

Notes: standard deviations in parentheses

2.3 Estimating the Effect of Education on Earnings

The reduced form equation relating earnings to educational attainment, based on the human capital earnings function of Mincer (1974), typically takes the form

$$\log y_i = \beta_0 + \beta_1 A_i + \beta_2 A_i^2 + \beta_3 e_i + \epsilon_i \quad (2.1)$$

where y is earnings, A is age which proxies for workforce experience, e is years of schooling completed, and ϵ is a statistical residual. This is the workhorse model for estimating the effect of education on earnings. Underlying this general specification is the assumption that $\log y \sim \mathcal{N}(\mu, \sigma_\epsilon^2)$, with $\mu = \beta_0 + \beta_1 A + \beta_2 A^2 + \beta_3 e$ and $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$. Because both y_i and e_i are determined in part by unobserved ability, $E(\epsilon|e) \neq 0$. Therefore, OLS estimation of (2.1) is biased. If suitable instruments can be found for e_i , β_3 can be interpreted as the average causal effect of education on earnings.³

In this section our aim is to estimate (2.1), or a similar equation, in a way that allows us to predict the future earnings of school-aged children, as plausibly forecast by parents in 1991-1994.⁴ We will estimate (2.1) separately on KHDS cross-sections from 1991, 2004, and 2011. In order to do this we

³The literature on estimation of (2.1) and similar specifications of the joint earnings-experience-education distribution is voluminous. See Card (1999) and Carneiro *et al* (2010) for recent treatments that also discuss the literature in detail.

⁴Results are not provided for the wide variety of models discussed in this section but not ultimately used in the paper. All results available upon request.

must decide how the education variable e should enter (2.1), balancing the paper's aims and data limitations against the need for consistent estimation. In the development literature, dummy variables for completion of primary and secondary school are often used in place of the single schooling variable e . This is both because sample sizes tend to be smaller in developing country data sets, and because the signaling and/or sheepskin effects from completion of schooling phases are considered particularly important in countries where government jobs with minimum educational requirements figure prominently in the formal labor market, and where the amount of learning and skills acquisition that takes place in schools is questionable. Because such discontinuities may very well be present in Tanzania, we could replace e_i in (2.1) with the pair of dummy variables (P_i, S_i) that take a value of 1 if an individual completes primary or secondary school, respectively, and 0 otherwise:

$$\log y_i = \beta_0 + \beta_1 A_i + \beta_2 A_i^2 + \beta_4 P_i + \beta_5 S_i + \epsilon_i \quad (2.2)$$

There are numerous drawbacks to this approach. The first is that under (2.2), a marginal year of education only increases earnings if it represents completion of primary or secondary school. While this may be true in the labor market, such a specification does not recognize that at the time the annual enrollment decision is made, the value of a non-terminal year of school is the option value it holds for eventual completion of a given schooling phase.

If we were going to model the marginal enrollment decision - the decision to acquire another year of schooling - in a structural setting, then option value would naturally be built into the model. But we will estimate the schooling choice decision in reduced form, and there is no way to derive an option value for non-graduation schooling years from (2.2) without adding numerous assumptions into what is already a heavily parameterized model.

One possible remedy is to include both e_i and (P_i, S_i) in the specification. We estimated a variety of such models, however, and in nearly all cases the positive effects of education are subsumed in the dummy variables, and the coefficient on e is negative. This is probably driven by the small but not inconsequential number of respondents who quit school just before completion of primary or secondary school. These individuals may have suffered a negative shock that both interrupted schooling and reduced future earnings.⁵ Whatever the correct interpretation, it is likely that the negative coefficient on e is a result of omitted variable bias, and does not represent the true effect of education on earnings in non-graduation years.

We could circumvent this shortcoming if we could simultaneously instrument for (e_i, P_i, S_i) . However, for reasons discussed momentarily, we only have one reliable instrument: father's education, f_i . While f_i performs well as an instrument for e_i alone, it is a very poor predictor of P_i and S_i . Functions of f_i , such as dummy variables for father's completion of primary and

⁵Or causality may run in the other direction: the shock could itself be a reduction in household earnings due to catastrophic loss of health or assets, which is transmitted to the next generation.

secondary school, fare no better. The instrument seems to do well on average, but not around the thresholds for completion of primary and secondary school.

A simpler way to allow for both marginal effects of non-graduation years and discontinuities at completion of primary and secondary school would be to include a full suite of dummy variables, one for each year of completed education. The marginal effect of each year of schooling would then be estimated as the mean effect, within education years, conditional on experience and its square. Again, because of small sample sizes in some education-year cells, this technique produced unreliable estimates in all of the estimated cross-sections.

Lastly, it may be possible to improve on (2.1) by including a second or third degree polynomial in education on the right hand side. Unfortunately, this approach tended to produce negative marginal returns in some (or many) educational years, so it too was abandoned.

Given all of the above shortcomings, and noting that for studying the schooling choice problem it is critical that the marginal value of schooling reflect to some degree both the causal effect of education on earnings, if there is one, and the option value for future continuation of schooling, we opt for the simple linear specification of (2.1), with the addition of a gender dummy and a gender-education interaction term:

$$\log y_i = \beta_0 + \beta_1 A_i + \beta_2 A_i^2 + \beta_3 e_i + \beta_4 M_i + \beta_5 M_i e_i + \epsilon_i \quad (2.3)$$

where $M_i = 1$ if person i is male, and 0 otherwise. If we can consistently estimate (2.3), then β_3 gives the average effect of a marginal year of schooling on women's earnings, and $\beta_3 + \beta_5$ gives the average effect of a marginal year of schooling on men's earnings.

Of course, if we use OLS to estimate (2.3), then the usual omitted variable problem biases our estimates of β_3 and β_5 downwards. We need to instrument for e_i . Unfortunately, there are very few instruments available in all survey rounds that are exogenous to earnings but not to education. Parents' levels of educational attainment are the only (arguably) reliable instruments that were gathered in 1991, 2004 and 2010, and only father's education was observed for a large enough number of respondents. Father's education is only a reliable instrument under the assumption that it does not determine child's earnings directly, e.g. through family connections or better farm management prior to inheritance, but only indirectly via the household taste for or access to education. If we were to restrict the 2004 and 2010 cross-sections to individuals who were school-aged during 1991-1994 - i.e., to the population of interest for the school choice problem - we could use various characteristics from individuals' childhood homes and communities to instrument for education.⁶ However, we lose nearly 75% of the 2004 and 2010 sample if we remove individuals not interviewed in 1991-1994.⁷ Also, we cannot use

⁶For example, we could make use of supply side schooling characteristics, as has been done in numerous papers.

⁷This is because a very costly and substantial effort was made in 2004 to interview all of the original KHDS respondents and all members of all of the households that they had formed or joined in the intervening decade. The latter group turned out to be significantly

supply-side or childhood IVs in the 1991 earnings-education estimation, so we would end up with different specifications in different cross-sections, making the predicted earnings incomparable. We are therefore left with only f_i as an instrument for e_i .

Unfortunately, f_i is also the only variable not already in (2.3) that is both known at the time of educational investments in 1991-1994 and gathered, via recall, from everyone interviewed in every round of KHDS. Because we want to estimate *subjective* returns to education using information known at the time of schooling decisions, it would be advantageous to add as many variables as possible to (2.3), to increase the degree of subjective variation in earnings forecasts. This is an argument for including f_i directly in the earnings equation, rather than using it as an instrument. We therefore do both, for sake of comparison. One interpretation of specifications that include f_i directly and are estimated via OLS is that father's education proxies for unobserved ability. If we believe that inclusion of f_i solves the omitted variable problem, the estimated coefficient vector is unbiased and consistent.

2.3.1 Results: earnings and education

In Table 2.2 we show the results of various specifications of the earnings equation, estimated separately for each cross section. The dependent variable in all regressions is average monthly individual earnings in 2010 Tanzania shillings. Within each year, the first column shows the results of OLS es-

larger than the former.

timination of equation (2.3). The second column shows the same regression, limited to respondents for whom father's education is available. The results in columns 1 and 2 are largely similar both within and across survey years; a marginal year of education is associated with an 8-10% increase in earnings, except for men in 1991 for whom the increase is 4%. Column 3 shows OLS results with f_i and $f_i \times M_i$ included on the right hand side. In 1991 and 2004 the sum of the coefficients on these two terms is essentially zero. If father's education is indeed a reliable proxy for unobserved ability, this suggests that for men in the sample, ability is exogenous to either earnings or education (or both). More likely is that f_i picks up the effects of numerous determinants of education. Column 4 shows the IV results, with f_i instrumenting for e_i . The IV results place the return to a marginal year of education in the neighborhood of 16-17% for women and 9-16% for men.

The estimated returns, though very high, are not unprecedented in the literature from developing countries (Psacharapolous and Patrinos, 2002). The marginal year results, as well as the returns implied by annual compounding of 4 and 8 marginal years of education,⁸ are shown in Table 2.3.⁹ Only the IV results and the OLS results from column 3 are shown. The IV results for women imply that conditional on experience, female primary school gradu-

⁸In the data, primary and secondary school graduates have on average 4 and 8 more years of education, respectively, than those who do not complete primary school.

⁹The results for "Primary" and "Secondary" completion in Table 2.3 are not directly comparable to the results one gets by using dummy variables for completing these phases of education, because the former rely on the conditional expectation across all education years, while the latter compare the conditional expectation between education subgroups.

Table 2.2: Estimation results: Mincerian returns functions in 1991, 2004 and 2010

	1991				2004				2010			
	OLS 1	OLS 2	OLS 3	IV	OLS 1	OLS 2	OLS 3	IV	OLS 1	OLS 2	OLS 3	IV
Education (a)	0.108 0.01***	0.097 0.01***	0.082 0.01***	0.165 0.02***	0.093 0.01***	0.088 0.01***	0.076 0.01***	0.18 0.02***	0.092 0.02***	0.087 0.02**	0.076 0.02**	0.163 0.03***
Education x Male (b)	-0.071 0.01***	-0.05 0.01***	-0.035 0.02*	-0.072 0.03*	-0.006 0.01	-0.011 0.01	0.001 0.01	-0.069 0.03**	0.005 0.01	0.005 0.01	0.011 0.01	-0.003 0.02
Age	0.141 0.01***	0.116 0.01***	0.118 0.01***	0.123 0.01***	0.143 0.00***	0.127 0.00***	0.13 0.00***	0.121 0.00***	0.118 0.00***	0.09 0.01***	0.094 0.00***	0.084 0.01***
Age ² / 1000	-1.27 0.08***	-1.055 0.08***	-1.074 0.08***	-1.079 0.08***	-1.265 0.04***	-1.123 0.05***	-1.143 0.05***	-1.003 0.06***	-1.032 0.01***	-0.743 0.04***	-0.768 0.03***	-0.624 0.05***
Male	1.279 0.09***	1.32 0.09***	1.34 0.09***	1.311 0.16***	0.973 0.06***	1.062 0.07***	1.138 0.07***	1.307 0.15***	1.086 0.07***	1.121 0.07***	1.156 0.08***	1.095 0.14***
Father's educ			0.034 0.01**				0.034 0.01***				0.034 0.00***	
Father's educ x Male			-0.031 0.02*				-0.032 0.01***				-0.015 0.01*	
R ²	0.394	0.38	0.382	0.363	0.426	0.401	0.404	0.372	0.437	0.385	0.389	0.352
N	2466	2007	2007	2007	7005	5041	5041	5041	7157	3699	3699	3699
F-stat: (a) + (b) = 0	11.272	19.946	16.509	6.311 ^a	128.14	80.658	83.387	14.27 ^a	12.782	9.485	10.074	13.515 ^a

Note: * sig at 5%, ** sig at 1%, *** sig at 0.1%; standard errors clustered at region level in 2010, village level in 1991 and 2004; education measured as years completed; constant not shown; dependent variable is log of earnings by persons aged 15+ and not currently in school; ^aChi-square statistic

Table 2.3: Marginal and threshold returns, calculated from Table 2.2

	Women			Men		
	Marginal year	Primary	Secondary	Marginal year	Primary	Secondary
1991 OLS	8.2%	37.1%	87.9%	4.7%	20.2%	44.4%
1991 IV	16.5%	84.2%	239.3%	9.3%	42.7%	103.7%
2004 OLS	7.6%	34.0%	79.7%	7.5%	33.5%	78.3%
2004 IV	18.0%	93.9%	275.9%	11.1%	52.4%	132.1%
2010 OLS	7.6%	34.0%	79.7%	6.5%	28.6%	65.5%
2010 IV	16.3%	82.9%	234.7%	16.0%	81.1%	227.8%

Notes: "Marginal year" columns are estimated coefficients; primary and secondary returns are the annually compounded return from 4 and 8 years of education, respectively

ates earn 85-90% more than women with only some primary education. On average, secondary school graduates earn nearly 2.5 times as much as those who do not finish primary. The comparable figures for men are lower, and are less stationary across the three survey years, but the estimated expected returns are nonetheless very substantial.

It is worth noting that after 1991, the gender-education term is insignificant in all specifications except the 2004 IV. This, however, reflects an *increase* in the gender difference in returns across the years of the KHDS. In 1991, men with no education earn approximately 2.3 times as much as women with no education, but that gap narrows with each successive year of schooling. In specifications from later years, except the 2004 IV, men earn 2-2.1 times as much as women, regardless of education. The key implication of these findings is that across cross-sectional years and specifications, the returns to education are especially high for girls.

If there is a trend in the returns results, it is that for women the returns to

education have been and remain high across the 20 years of KHDS, while the returns for men have increased steadily to match the figures for women. The female earnings-experience-education distribution is more or less stationary. This suggests that, at least with regard to expected log earnings, the predicted marginal effect of education is invariant to the choice of cross-sectional results used to make the prediction. For men, steady increases in the returns to education from 1991-2010 suggest that educational choices made in 1991-1994, conditional on the contemporaneous returns distribution, would have been sub-optimal.

2.3.2 Estimating the variance of log earnings

The results of the previous subsection can be used to construct predictions of expected log earnings conditional on a marginal educational choice, $E(\log y|X, \hat{\beta}, e)$, for children who are school-aged in 1991-1994. Let $\hat{\mu} = X\hat{\beta}$ denote predicted expected log earnings. We also need a predictor of the subjective variance of log earnings, $\hat{\sigma}_\epsilon^2$, in order to fully characterize the distribution in levels.¹⁰ Following Hildreth and Houck (1968) and Just and Pope (1978), we calculate $r_i^2 = \log y_i - X_i\hat{\beta}$, where X_i and $\hat{\beta}_i$ are the design matrix and estimated coefficient vector from one of the regressions in the previous section, respectively, and then run regressions based on the following specification:

¹⁰If $y \sim \log -N(m, s^2)$, then (μ, σ_ϵ^2) such that $\log y \sim N(\mu, \sigma^2)$ are sufficient statistics for (m, s^2) .

$$r_i^2 = \gamma_0 + \gamma_1 A_i + \gamma_2 A_i^2 + \gamma_3 e_i + \gamma_4 M_i + \gamma_5 M_i e_i + \nu_i \quad (2.4)$$

That is, for each of the regressions of $\log y$ on individual characteristics reported in the previous section, we square the residuals and regress them back on the same set of independent variables (or instruments, depending on the specification). This is a two-step method for modeling conditional heteroskedasticity. The very premise under which we estimate (2.4) - the premise that the variance of the stochastic component of log earnings is not constant across individuals - implies that the results of the previous subsection are inefficient, though unbiased. However, while we take some interest in the returns equations themselves, we are not particularly concerned with hypothesis testing in this first estimation stage, so we accept the inefficiency in the log earnings regressions as the cost of explicitly modeling both the expectation and the variance of earnings.

Table 2.4 shows the results of OLS and IV estimation of (2.4), using the squared residuals from the corresponding Mincerian returns estimate. The explanatory powers of the models are low, suggesting that the variance of the stochastic component of log earnings is largely orthogonal to individual characteristics. Education does, however, have a statistically significant effect on the variance of log earnings in many specifications. For women in 1991 and 2004, increases in education are associated with a small but statistically significant decrease in the variance of log earnings. For men, the

IV results in 1991 and 2010 suggest that education is associated with an increase in the variance of earnings.¹¹ Like the gender difference in the log earnings equations, this gender difference is not surprising. Anecdotal and observational evidence suggests that labor markets for men and women in Tanzania are characterized by substantial gender discrimination, though less so in 2010 than in 1991. It appears that education has a stabilizing effect on female earnings, likely by providing access to waged formal sector jobs. For men the effect is the opposite, though this may be due to a greater number of men seeking secondary and advanced education without a corresponding increase in demand from employers, leaving numerous well-educated men with second-best employment on farms or in low pay manual labor.

¹¹We use the IV specification to estimate (2.4) because the consistency results in Just and Pope (1978) require that we use the same set of explanatory variables in the first stage and in the squared residual regression. However, if preferences for stable waged income, relative to the higher variance income observed from farming and self-employment in Tanzania, are correlated with educational attainment, then the IV is justified on its own merits.

Table 2.4: Estimation results with squared residuals from Table 2.2 as dependent variable

	1991				2004				2010			
	OLS 1	OLS 2	OLS 3	IV	OLS 1	OLS 2	OLS 3	IV	OLS 1	OLS 2	OLS 3	IV
Education (a)	-0.094 0.02***	-0.087 0.02***	-0.067 0.02**	-0.107 0.04*	-0.05 0.01**	-0.051 0.02*	-0.05 0.02*	-0.036 0.04	0.009 0.01	0.011 0.01	0.02 0.01	-0.023 0.02
Education x Male (b)	0.08 0.02***	0.069 0.03*	0.026 0.03	0.251 0.05***	-0.025 0.04	-0.029 0.05	-0.04 0.05	0.087 0.11	-0.043 0.01***	0.031 0.02	0.003 0.02	0.175 0.04***
Age	-0.031 0.01*	-0.02 0.01	-0.012 0.01	-0.021 0.01	-0.05 0.01***	-0.035 0.02*	-0.03 0.02	-0.044 0.02**	-0.066 0.01***	-0.047 0.01***	-0.043 0.01***	-0.052 0.01***
Age ² / 1000	0.389 0.14**	0.274 0.14	0.218 0.14	0.312 0.14*	0.673 0.12***	0.533 0.14***	0.504 0.14***	0.632 0.15***	0.933 0.09***	0.677 0.07***	0.652 0.07***	0.725 0.06***
Male	-0.404 0.16*	-0.408 0.18*	-0.452 0.18*	-1.342 0.28***	0.461 0.27	0.483 0.35	0.429 0.38	-0.263 0.68	0.421 0.05***	-0.212 0.07*	-0.29 0.07**	-1.08 0.13***
Father's educ			-0.017			0.011				-0.01		
Father's educ x Male			0.01			0.01				0.01		
			0.096			0.03				0.065		
			0.02***			0.02				0.00***		
R ²	0.03	0.027	0.033	.	0.015	0.013	0.013	0.006	0.016	0.01	0.011	0.003
N	2466	2007	2007	2007	7005	5041	5041	5041	7157	3699	3699	3699

Note: * sig at 1%, ** sig at 5%, *** sig at 10%; standard errors clustered at region level in 2010, village level in 1991 and 2004; education measured as years completed, constant not shown; dependent variable is square of residual from first stage earnings equation

2.3.3 Predicting the expectation and variance of earnings in levels

In this final sub-section our aim is to recover estimates of the first two subjective, conditional moments of the level earnings distribution, $\hat{y}_i = E(y_i|x_i, \hat{\beta}, \hat{\gamma})$ and $\hat{\sigma}_i^2 = Var(y_i|x_i, \hat{\beta}, \hat{\gamma})$, where x_i is the vector of explanatory variables used in (2.3) and (2.4), and $(\hat{\beta}, \hat{\gamma})$ are the corresponding coefficient vectors. We use only the results from columns 3 and 4 within each cross-sectional year, i.e. only the OLS results with father's education proxying for ability, and the IV results with father's education instrumenting for education.

There is a small literature on the recovery of $E(y_i|x_i, \hat{\beta})$ from the log-normal regression $\log y_i = x_i\beta + \epsilon$, $\epsilon \sim N(0, \sigma_\epsilon^2)$.¹² It is well known that the naive, backwards transform $\hat{y}_i = \exp(x_i\hat{\beta})$ is biased downwards. Bradu and Mundlak (1970) derive the uniformly minimum variance unbiased (UMVU) estimator of $E(y|x, \hat{\beta})$, conditional on the estimated parameters of the log-linear distribution. El-shaarawi and Viveros (1997) derive an alternative estimator that corrects for bias in the first stage restricted maximum likelihood estimation of the log-linear mean ($x\hat{\beta}$) and variance ($\hat{\sigma}^2$). More recently, Shen and Zhu (2008) develop two estimators of $E(y_i|x_i, \hat{\beta})$, one that minimizes bias and the other that minimizes mean-square error. All of these

¹²The papers cited in this paragraph are concerned exclusively with unbiased estimation of the mean $E(y_i|x_i, \hat{\beta})$. To my knowledge there is no work specifically on unbiased recovery of $Var(y_i|x_i, \hat{\beta})$.

estimators depend on $x_i\hat{\beta}$ and $\hat{\sigma}_\epsilon^2$, the expectation and variance of ϵ from the log-linear stage, which are sufficient statistics for \hat{y}_i and $\hat{\sigma}_i^2$. These estimators, however, are derived with an eye toward unbiased estimation of only the expectation \hat{y}_i , not $(\hat{y}_i, \hat{\sigma}_i^2)$ together. Furthermore, while a subjective component to the variance is built naturally into each estimator through the term $x_i'(X'X)^{-1}x_i$, which enters $\hat{\sigma}_\epsilon^2$, such estimators cannot defensibly be used to make out-of-sample predictions, which we will do for children in 1991-1994 who were not re-interviewed in later rounds of KHDS. Therefore, rather than rely on these estimators, we make direct use of our linear predictions for the expectation and variance of log earnings, $x_i\hat{\beta}$ and $x_i\hat{\gamma}$, and calculate:

$$\hat{y}_i = \exp(x_i\hat{\beta} + \frac{x_i\hat{\gamma}}{2}) \quad (2.5)$$

$$\hat{\sigma}_i^2 = [\exp(x_i\hat{\gamma}) - 1] \exp(2x_i\hat{\beta} + x_i\hat{\gamma}) \quad (2.6)$$

Estimators (2.5) and (2.6) are derived directly from the definition of the log-normal distribution. Equation (2.5) gives an unbiased predictor of the subjective expectation of y_i if the estimators $\hat{\beta}$ and $\hat{\gamma}$ are unbiased (Shen and Zhu, 2008).¹³ To my knowledge, the efficiency properties of these estimators under the assumption of conditional heteroskedasticity in the log earnings

¹³Because we have not explicitly modeled the variance of the stochastic component of the variance equation, σ_ν^2 , it is likely that the variance prediction is biased slightly downwards. To attempt a fix for this bias here would take us well beyond the scope of the paper. However, because the bias suppresses variation in $\hat{\sigma}_i^2$, any significant effects of the predicted second moment on enrollment decisions can be considered lower bounds on the true effect.

equations are not known.¹⁴

Using (2.5) and (2.6), we build subjective distributions of log-normal earnings, evaluated at the future age of 30 years, conditioning on the characteristics of school-aged children in the early 1990s. Recall that with age fixed, these earnings distributions are subjective only up to gender, father's education, and current grade (or marginal grade, if not enrolled in school).

Table 2.5 displays correlation coefficients of the predicted moments for the OLS results with father's education and the IV results, across the three cross-section years. For a given estimation method (OLS or IV), the correlations between expected earnings in the three survey years are all greater than 0.9. This suggests that the predictable component of the expectation of the earnings distribution is essentially stationary from 1991 to 2010. However, the correlation in expected earnings across estimation methods and years taken together is lower, ranging between 0.6 and 0.92. Thus, the degree to which we believe father's education is better used as an instrument for education rather than a proxy for unobserved ability has some bearing on the predicted first moment. The inter-year correlations of the predicted variance of earnings are uniformly lower, except for the IV results which are highly correlated. The inter-method and within-OLS correlation coefficients range between 0.1 and 0.77. This implies that to the degree that the predictable component of the variance of future earnings factors into educational

¹⁴What is known is that (2.5) is not the same as the minimum variance estimator of y_i under the assumption of homoskedasticity; this would be the UMVU of Bradu and Mundlak (1970).

Table 2.5: Correlation coefficients of expectation and variance estimates

	A. Estimated E[earnings] at age 30						B. Estimated Var[earnings] at age 30					
	1991 OLS	2004 OLS	2010 OLS	1991 IV	2004 IV	2010 IV	1991 OLS	2004 OLS	2010 OLS	1991 IV	2004 IV	2010 IV
1991 OLS	1						1					
2004 OLS	0.99	1					0.91	1				
2010 OLS	0.93	0.93	1				0.68	0.49	1			
1991 IV	0.70	0.74	0.90	1			0.19	0.10	0.68	1		
2004 IV	0.76	0.80	0.92	0.99	1		0.30	0.23	0.77	0.94	1	
2010 IV	0.60	0.64	0.84	0.98	0.95	1	0.13	0.05	0.56	0.97	0.83	1

investment decisions, better schooling choices¹⁵ will be made by those who anticipate the future distribution of returns.

Figures 2.1 and 2.2 depict the variation in the predicted moments. Figure 1 shows 4 log-normal distributions for each year and estimation method: the earnings distributions associated with the minimum and maximum predicted \hat{y}_i distributions (with their associated $\hat{\sigma}_i^2$), the distribution formed by the average \hat{y}_i and $\hat{\sigma}_i^2$, and the observed distribution of earnings among 30 year-olds in 2010, for sake of comparison. The lack of variation across years and estimation methods is evident, although some underlying variation is not viewable in a graph of only the minimum, maximum and mean.

What is most striking in Figure 2.1 is the substantial degree of variation *between* the predicted distributions, driven by gender, father's education and marginal year of education. Clearly it makes little sense to speak of a single marginal return to education, when the stock of education and other child characteristics critically affect the subjective distribution of returns. If we were able to condition on more of the observed heterogeneity in child charac-

¹⁵“Better” in the sense that they will have higher expected utility *ex ante*.

teristics when we fit the returns distributions, and if we allowed for essentially zero return to some of the middle years of primary and secondary school by using categorical dummies for educational phases rather than the linear educational term,¹⁶ the variation in subjective returns distributions would likely be even more striking.

The increase in variance associated with higher levels of education is partly mechanical, as the variance of earnings increases exponentially with the mean. To adjust for this, Figure 2.2 shows the average coefficient of variation from the OLS and IV estimates for each of the three cross-sectional years. The precipitous increase at grade 13 can be ignored, as it is caused by the small number of enrollees at that grade. There is no clear pattern in Figure 2.2. After normalization, it appears that the average subjective variance based on the 1991 and 2004 OLS results falls with grade, while the variance based on the 2010 IV results increases with the grade. However, the most we can conclude from Figure 2.2 is that the increase in expected value and the increase in variance as education increases are largely off-setting. The effect this has on enrollment decisions will depend on parents' and students' attitudes toward risk.

¹⁶Results using dummy variables for various educational achievement categories are shown in an Appendix.

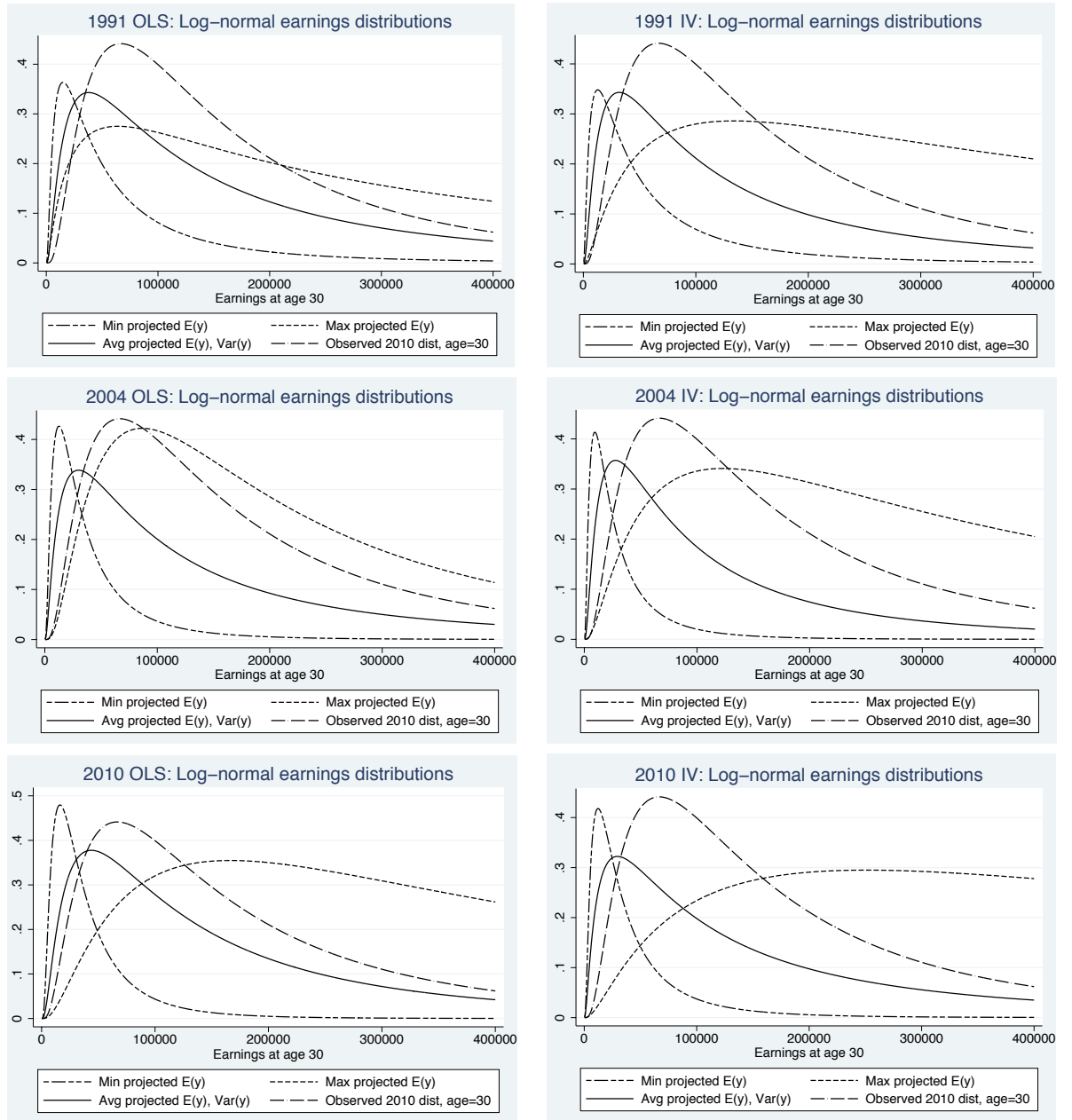


Figure 2.1: Subjective earnings distributions

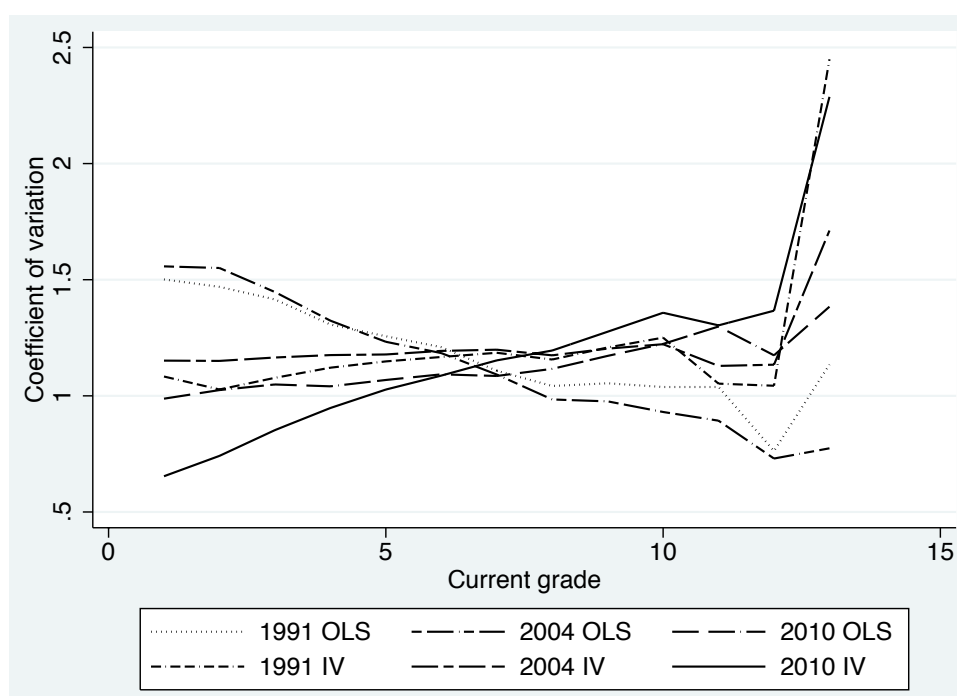


Figure 2.2: Average coefficient of variation from subjective distributions

2.4 Enrollment Decisions

We now turn to the educational investment decisions made in the early 1990s. Because of overlap in interview years, we observe a maximum of 3 years of school enrollment (or non-enrollment) for each school-aged child. Although the panel is short, it does allow us to control for unobserved individual effects in school choices.¹⁷ This bypasses most of the identification concerns in Glewwe and Kremer (2005), who emphasize the confounding effects of unobserved individual ability in most studies of school choice in developing countries. In addition to controlling for unobserved ability, we also control for school costs by including the median total school fees within district-gender-grade subgroups as a regressor. And of course, we control for the subjective return to the marginal year of education by conditioning on \hat{y}_i and $\hat{\sigma}_i^2$ from the previous section.

It is well known that conditional fixed effects estimation of a binary response model with panel data is not feasible, because of the incidental parameters problem, and that unconditional fixed effects estimation in a panel is biased. Instead, we posit a random effects probit model. Let α_i be the time invariant, unobserved individual effect, \hat{W}_{it} be the matrix of enrollment determinants, including the predicted expectation and variance of earnings, c_{it} be the outcome dummy, which takes a value of 1 if individual i is enrolled in school in period t and 0 otherwise, and Φ be the cumulative standard nor-

¹⁷Because the panel is so short, we also make no attempt to correct for the unbalanced nature of the data.

mal distribution. The key assumptions of the random effects probit model are:

$$P(c_{it} = 1 | \hat{W}_i, \alpha_i) = P(c_{it} = 1 | W_{it}, \alpha_i) \quad \text{for } t = 1 \dots T \quad (2.7)$$

$$P(c_{it} = 1 | \hat{W}_{it}, \alpha_i) = \Phi(W_{it}\delta + \alpha_i) \quad \text{for } t = 1 \dots T \quad (2.8)$$

$$c_{i1} \dots c_{iT} \text{ independent conditional on } (W_i, \alpha_i) \quad (2.9)$$

$$\alpha_i | \hat{W}_i \sim \mathbb{N}(0, \sigma_\alpha^2) \quad (2.10)$$

In Assumption (2.7), the term \hat{W}_i contains $(\hat{W}_{i1} \dots \hat{W}_{iT})$ for each individual $i = 1 \dots N$. This assumption, in essence, states that \hat{W}_{it} is strictly exogenous conditional on the unobserved effect α_i . Assumption (2.8) is the usual probit assumption that the probability of a positive outcome is given by the value of the cumulative standard normal distribution, with the linear index term $\hat{W}_i\delta + \alpha_i$ as the argument. Assumption (2.9) states that the observed determinants \hat{W}_i and the unobserved α_i fully characterize the predictable component of the school choice decision, so that enrollment decisions within-person are independent after conditioning on (\hat{W}_i, α_i) . The last assumption is the strongest of the four, asserting both that the α_i and \hat{W}_i are independent and that the unobserved effects are normally distributed. Independence of \hat{W}_i and α_i is necessary to identify the random effects probit

model.¹⁸

Under assumptions (2.7) - (2.10), we maximize the conditional likelihood of $(\delta, \sigma_\alpha^2)$ after integrating out α :

$$\mathcal{L}(\delta, \sigma_\alpha^2 | x_i, c_i) = \int_{-\infty}^{\infty} \left[\prod_{t=1}^T \Phi(\hat{W}_{it}\delta + \alpha_i)^{c_{it}} [1 - \Phi(\hat{W}_{it}\delta + \alpha_i)]^{1-c_{it}} \right] \left(\frac{1}{\sigma_\alpha} \right) \phi\left(\frac{\alpha_i}{\sigma_\alpha}\right) d\alpha_i \quad (2.11)$$

Taking the log of (2.11) and multiplying across individuals gives the conditional log likelihood of $(\delta, \sigma_\alpha^2)$. Although sample sizes are not small for a developing country survey, we bootstrap the standard errors to mitigate the effects of sampling error, small-sample biases, and the presence of the predicted regressors in \hat{W}_{it} .

2.4.1 Results of random effects probit model

Tables 2.6 and 2.7 display the coefficients and marginal effects from the estimation of (2.11). Results are reported separately for values of \hat{y}_i and $\hat{\sigma}_i^2$ constructed from each cross-sectional year (1991, 2004 and 2010) and from each method (OLS and IV). Repeat conditioning on father's education and gender in the random effects probit estimation, after their inclusion in the first-stage estimates of the expectation and variance of earnings, is not problematic, given that the first-stage predicted values are determined by at

¹⁸Alternative specifications are possible, if we are willing to parameterize the relationship between c_i and \hat{W}_i . See Chamberlain (1980), or Wooldridge (2002) for a discussion.

Table 2.6: Coefficients from random effects probit models

	E(y) and Var(y) from OLS			E(y) and Var(y) from IV		
	1991	2004	2010	1991	2004	2010
Male	-3.399	-3.334	-1.427	-0.241	-0.722	0.209
	0.86***	0.91***	0.38***	0.26	0.28***	0.19
Age	-0.227	-0.245	-0.226	-0.213	-0.239	-0.187
	0.03***	0.03***	0.02***	0.02***	0.03***	0.03***
Asset index	0.169	0.167	0.171	0.177	0.159	0.193
	0.07**	0.06***	0.07**	0.07***	0.06***	0.08**
Birth order	0.036	0.029	0.021	0.021	0.026	0.019
	0.04	0.05	0.04	0.05	0.03	0.04
School fees (x 10e-3)	-0.122	-0.129	-0.118	-0.119	-0.123	-0.12
	0.02***	0.02***	0.02***	0.02***	0.02***	0.02***
Father's education	-0.014	-0.003	0.014	0.066	0.06	0.07
	0.04	0.03	0.02	0.02***	0.02***	0.02***
E(y) at age 30 (x 10e-6)	66.978	77.152	31.041	16.307	33.492	5.875
	12.87***	12.73***	7.77***	6.13***	8.74***	3.17*
Var(y) at age 30 (x 10e-12)	-39.185	-50.535	-24.451	-4.076	-30.028	0.215
	10.42***	15.53***	12.69*	7.17	16.29*	3.24
District effects	Yes	Yes	Yes	Yes	Yes	Yes
N	2592	2592	2592	2592	2592	2592
Prediction accuracy	67.5%	67.7%	67.5%	67.3%	67.6%	66.9%

Note: * sig at 10%, ** sig at 5%, *** sig at 1%; random effects probit model with bootstrapped standard errors; dependent variable is dummy for school enrollment; district effects and constant not shown

least one excluded variable (age = 30) and passed through highly non-linear transforms, (2.5) and (2.6).

Most coefficients in Table 2.6 have the expected sign. Younger students are more likely to be enrolled in school; conditional on age, birth order does not have a significant effect on the probability of enrollment. As we would expect, the coefficient on father's education is insignificant when used as a proxy for ability in the first stage, but positive and significant when used as

Table 2.7: Marginal effects from random effects probit models

	E(y) and Var(y) from OLS			E(y) and Var(y) from IV		
	1991	2004	2010	1991	2004	2010
Male	-0.902	-0.896	-0.516	-0.094	-0.277	0.082
	0.094***	0.106***	0.116***	0.13	0.108**	0.073
Age	-0.09	-0.097	-0.089	-0.084	-0.094	-0.074
	0.011***	0.009***	0.012***	0.013***	0.013***	0.010***
Asset index	0.067	0.066	0.068	0.07	0.063	0.077
	0.024***	0.022***	0.022***	0.029**	0.029**	0.026***
Birth order	0.011	0.009	0.006	0.006	0.008	0.005
	0.02	0.018	0.017	0.016	0.02	0.013
School fees (x 10e-3)	-0.048	-0.051	-0.046	-0.047	-0.049	-0.047
	0.008***	0.008***	0.006***	0.007***	0.007***	0.007***
Father's education	-0.006	-0.001	0.006	0.026	0.024	0.028
	0.016	0.018	0.01	0.010**	0.008***	0.009***
E(y) at age 30 (x10e-6)	26.311	30.354	12.208	6.414	13.167	2.309
	6.028***	5.622***	2.911***	3.160**	4.113***	1.390*
Var(y) at age 30 (x10e-12)	-15.359	-19.83	-9.604	-1.602	-11.795	0.085
	4.371***	6.000***	5.429*	4.202	9.351	1.793
District effects	Yes	Yes	Yes	Yes	Yes	Yes
N	2592	2592	2592	2592	2592	2592

Note: * sig at 10%, ** sig at 5%, *** sig at 1%; marginal effects from random effects probit model with bootstrapped standard errors; dependent variable is dummy for school enrollment; district effects and constant not shown

an instrument for education in the first stage. Higher school fees reduce the likelihood of enrollment. Most interestingly for this paper, the coefficients on the expectation and variance of earnings at age 30 are highly significant in most specifications, and they have the expected signs. It is noteworthy that the explanatory power of the model is invariant to the year and the first-stage estimation method. Thus, despite the observed variation in the moments of subjective returns distributions across years, we have no evidence to indicate whether households are more likely to make enrollment decisions using the forecasted distributions from 2004 and 2010 or the contemporaneous distribution from 1991.

Table 2.7 shows the marginal effects from the coefficients in Table 2.6, evaluated at the mean of each variable. To put the results in context: the predicted values of \hat{y}_i from the 1991 OLS results range from 27,347 to 183,766 TSH/month, and an increase in 1,000 TSH expected earnings per month leads to a 2.6% increase in the probability of enrollment. Correspondingly, $\hat{\sigma}_i^2$ from the 1991 OLS results ranges from $1,139.85 \times 10^6$ to $242,972.7 \times 10^6$, and an increase in variance of $1,000 \times 10^6$ TSH/month is associated with a 1.5% decrease in the probability of school enrollment. Results from other years and estimation methods are similar. Considering the substantial heterogeneity in predicted earnings distributions, these results have considerable economic significance. Furthermore, given that enrollment falls gradually as the marginal grade increases, these results cannot be driven by grade effects. Higher variance in earnings, conditional on expected earnings, school fees

and other characteristics, has an economically significant negative effect on enrollment. The opposite is true for higher expected earnings.

2.5 Conclusion

The results in this paper point to two general conclusions. First, while the economic returns to education did change to some degree over the study period, the changes were not significant enough (in the sense that they were too weakly correlated with conditional enrollment decisions) to have an effect on the predictability of school choice. Second, the expectation and the variance of the subjective earnings distributions have significant positive and negative effects, respectively, on the probability of school enrollment. This is in keeping with our initial hypothesis regarding investment under uncertainty by risk averse individuals.

These findings depend heavily on our use of particular parametric specifications for the returns distributions and the schooling choice process. Clearly it would be of interest to study problems of investment under uncertainty - such as educational enrollment decisions - with a firmer grasp on the role of subjective distributions of returns in school choice. Rather than explore other heavily parametric avenues, or resort purely to natural experiments and randomized control trials, one potentially fruitful direction for this line of research involves the measurement and analysis of subjective returns distributions. Gathering such data would allow us to further explore the effect

of heterogeneity in both realized and perceived conditional returns distributions on school choice.

2.6 Appendix

Calculation of individual-level consumption measure

Unfortunately, most of the questions that covered agricultural production and labor income in KHDS 1991-1994 and 2004 were removed from KHDS 2010, making it impossible to use earnings data from 2010 to fit an earnings-experience-education distribution to the data. Instead we use consumption (expenditure) data, which was collected in detail in all KHDS rounds. In developing countries, where home production of food is prevalent, consumption is closely correlated with the permanent component of household income (and is thus less volatile than income). In KHDS 1991-1994, consumption has a higher mean and a lower variance than income across all survey rounds. Most consumption data was gathered at the household level; we needed to assign consumption levels to individuals. In order to do so, we calculated the average share of household income, θ_{mas} , contributed by individuals of a particular gender, m , age group, a , and status in the household, s , in 2004:¹⁹

$$\theta_{mas} = \frac{\sum_{j=1}^{N_{2004}} \theta_{jh} \cdot \mathbb{I}(M_{jh} = m) \cdot \mathbb{I}(A_{jh} = a) \cdot \mathbb{I}(S_{jh} = s)}{\sum_{j=1}^{N_{2004}} \mathbb{I}(M_{jh} = m) \cdot \mathbb{I}(A_{jh} = a) \cdot \mathbb{I}(S_{jh} = s)} \quad (2.12)$$

where \mathbb{I} is the indicator function; N_{2004} is the 2004 cross-section sample size; θ_{jh} is the share of household income contributed by person j in house-

¹⁹See the next Appendix section for a discussion of the calculation of the 2004 income shares.

hold h . The triplet (M_{jh}, A_{jh}, S_{jh}) gives the gender, age group, and status in the household of person j in household h . Age group categories are: less than or equal to 15 years, 16-25 years, ..., 55-65 years, and 65+. Status in the household is divided into four categories: head, spouse of head, child of head, and other. We use only the 2004 cross-section to identify income shares because the individual-level earnings data in 2004 is more comprehensive than in 1991-1994, and because the tracking and interviewing of additional respondents in 2004 greatly increased the sample size. We use these shares to assign a proportion of household consumption to individuals in the 1991, 2004 and 2010 cross-sections, after re-normalizing to ensure that household consumption shares sum to 1. Note that this is a measure of the consumption share *generated*, not consumed, by each individual. The key assumption underlying this method is that the average proportion of income contributed by members of gender-age-status subgroup is stable from 1991-2010. Given that our interest is in the capacity for households to forecast education returns in the mid-term future, this assumption is not overly restrictive. We do not assume stationarity in the distribution of educational attainment by subgroup members, nor in the returns to education in particular sectors or for particular gender-age-status subgroups. Changes in these distributions will be picked up as changes in total household consumption, and assigned proportionally to individual household members. Some variation will be suppressed if the relative returns between subgroups changes substantially over the twenty study years, but there is nothing that can be done about this in

the absence of 2010 earnings data.

Estimating earnings in 2004

Because individual-level earnings is not observed for all rounds of KHDS, the relationship between earnings and consumption in 2004 plays a critical role in determining individual income measures for the three cross-sectional years of interest (see previous sub-section). In this section we discuss the construction of total individual returns in 2004.

Individual on-farm earnings are not directly observable. To assign earnings from own-farm agricultural labor to household members in the 2004 cross-section, we estimate a production function in logs and assign the estimated share of earnings to each family member. For robustness we run a variety of production specifications and compare the resulting earnings profiles. We include in these regressions all 2,079 households from the 2004 cross-section that list positive, non-trivial agricultural revenues. The general form of the estimated farm revenue production function, with h indexing households, is the following:

$$\log Y_h = \beta_1 \log K_h + \beta_2 \log A_h + \beta_3 \log L_h + \beta_4 Z_h + \epsilon_h \quad (2.13)$$

$$\iff y_h = \beta_1 k_h + \beta_2 a_h + \beta_3 l_h + \beta_4 z_h + \epsilon_h \quad (2.14)$$

where Y is the total value of agricultural output, K is the self-reported value of agricultural tools and equipment, A is the number of acres cultivated,

L is household labor, which is divided into male and female labor in most specifications, Z is other inputs such as Tropical Livestock Units (TLU) and the total value of purchased variable inputs such as seeds, labor, pesticides and transport, and ϵ is mean-zero error assumed to be uncorrelated with the independent variables. Lowercase variables are in logs. The estimated coefficient vector β assigns shares of agricultural earnings to the independent variables. Y includes the value of food consumed at home, as well as revenue from crop sales and from the sales of livestock products and processed crop products. Revenue from livestock sales is not included, as this is considered part of the household capital account. Labor is measured as the sum across all household members of monthly hours working with crops or livestock. In most specifications we allow for male and female labor (L_m and L_f , respectively) to enter separately. When combined, $L = L_m + L_f$. Tropical livestock units are measured in the standard fashion, with 1 TLU = 1 cattle = 10 sheep or goats.

Table 2.8 shows the results from a variety of specifications of the production function. Columns 1-4 include only households with positive values of all independent variables (ensuring that the log is defined). The specifications in columns 5-8 are the same as those in columns 1-4, with the exception that the level value of each independent variable is increased by 1, so that the log of all variables is defined for all households. The female labor share is substantially larger than the male share in the first 3 columns. Controlling for livestock holdings (column 3) eliminates the contribution of male labor,

most likely because limiting the sample to households with positive livestock holdings disproportionately weights the contribution of young boys tending livestock in relatively poor households. The explanatory power of the model is essentially constant across specifications. The row labeled “Sum of shares” gives the sum of the coefficients, excluding the constant. In order to fully apportion all revenue across the various inputs, we re-normalize the share coefficients by dividing each by the sum of shares. The adjusted shares of revenue for male, female and total labor (as applicable) are shown in the three rows below “Sum of shares”.

Using the shares from the 8 regressions shown in Table 2.8, the agricultural earnings of individual i in 2004, denoted y_{ai} , are given by:

$$\text{Gendered labor:} \quad y_{ai} = \left(M\alpha_{ihm}\beta_{L_m} + (1 - M)\alpha_{ihf}\beta_{L_f} + H\beta_K \right) Y \quad (2.15)$$

$$\text{Total labor:} \quad y_{ai} = \left(\alpha_{ih}\beta_L + H\beta_K \right) Y \quad (2.16)$$

where $M = 1$ for male workers and 0 otherwise; α_{ihj} for $j \in \{m, f\}$ is the percent of total male or female work hours, respectively, in household h , worked by individual i ; β_{L_j} for $j \in \{m, f\}$ is the coefficient on male or female labor, respectively, in model 1-3 or 5-7; β_L is the coefficient on total labor in model 4 or 8; $H = 1$ if the individual is the head of the household, 0 otherwise; β_K is the sum of non-labor coefficients in the regression, normalized by the sum of the shares; and Y is the total value of agricultural

Table 2.8: Agricultural revenue production shares in the 2004 cross-section

Model	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Male labor hours	0.049	0.054	-0.009		0.034	0.024	0.035	
	0.03	0.03	0.04		0.01***	0.01***	0.01***	
Female labor hours	0.112	0.09	0.068		0.03	0.022	0.032	
	0.03***	0.04*	0.05		0.01***	0.01**	0.01***	
Ag capital value	0.368	0.317	0.301	0.312	0.227	0.212	0.213	0.229
	0.03***	0.03***	0.04***	0.02***	0.01***	0.01***	0.01***	0.01***
Acres cultivated	0.26	0.248	0.185	0.285	0.244	0.234	0.239	0.252
	0.04***	0.04***	0.05***	0.03***	0.02***	0.02***	0.02***	0.02***
Total variable inputs		0.067				0.02		
		0.02***				0.00***		
TLU			0.143				0.085	
			0.03***				0.02***	
Total labor hours				0.186				0.058
				0.02***				0.01***
R ²	0.309	0.338	0.33	0.304	0.296	0.307	0.306	0.291
N	1039	851	519	1813	2079	2079	2079	2079
Sum of shares	0.789	0.776	0.697	0.783	0.535	0.512	0.604	0.539
Male L share (adj)	0.062	0.070	0		0.064	0.047	0.058	
Fem L share (adj)	0.142	0.116	0.098		0.056	0.043	0.053	
Total L share (adj)				0.238				0.108

Note: * sig at 5%, ** sig at 1%, *** sig at 0.1%; constant not reported; all variables in logs; dependent variable is log of agricultural output (TZS)

Table 2.9: Correlation between earnings, Models (1)-(8) from Table 2.8

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8
Model 1	1							
Model 2	0.99	1						
Model 3	0.91	0.85	1					
Model 4	0.86	0.89	0.67	1				
Model 5	0.88	0.94	0.61	0.88	1			
Model 6	0.89	0.95	0.64	0.88	1.00	1		
Model 7	0.89	0.94	0.63	0.88	1.00	1.00	1	
Model 8	0.86	0.89	0.67	1.00	0.88	0.88	0.88	1

output. This method of apportioning earnings attributes the return to all non-labor inputs to the head of the household, and divides the labor shares across household members in accordance with the contribution of each to total household agricultural labor. Table 2.9 shows the correlation in labor earnings - excluding the returns to other inputs that are allocated to the head - across the 8 specifications. All correlations not involving model 3, which assigned a share of 0 to male labor, are greater than 0.86.

Despite these high correlations, the choice of model still has important consequences for estimated earnings and subsequent regressions, because these high correlation coefficients mask significant level differences. Across models (1)-(8), the total value of agricultural earnings assigned to the head of the household ranges from 76 – 90%. Thus, in many households the labor earnings of a non-head will be twice as much under some specifications as under others. In general, the method is likely to overstate the head's earnings, since some household livestock or capital inputs are sure to be at least

jointly owned by other household members. For this reason we opted to use the results of model 1 to estimate earnings, as these have the highest labor share among models that allow male and female labor to enter separately.

As mentioned previously, we add the value of agricultural employment earnings in 2004, for which we have wage data, to these estimated on-farm individual earnings to arrive at total agricultural earnings for each individual in the 2004 cross-section.

Construction of the Asset Index

To control for household wealth, we estimate a single index of underlying wealth using a vector of observed asset holdings and household characteristics. This procedure, which has quickly become standard in empirical development economics, is explained in detail in Sahn and Stifel (2003) and Filmer and Pritchett (2001). We calculate both wave-specific indexes and a single index pooled across survey waves. We use the pooled measure in regressions, because the values of these indexes are unitless and therefore impossible to interpret in anything but relative terms. The asset index is constructed from the first principal factor underlying a vector that includes the number of various durable goods owned by the household and dummy variables indicating characteristics of the physical dwelling. Table 2.10 lists the within-wave and pooled means of the assets included in the index. Table 2.11 shows the factor loadings for all 5 asset indexes. All factor loadings in all waves have the expected sign. The value of most loadings is consistent

across waves, with the exception of the dummy variables for different types of lighting and some of the indicators for wall type, which vary across waves.

Table 2.10: Summary statistics for variables used in asset index, 1991-1994

	Wave 1		Wave 2		Wave 3		Wave 4		Pooled	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Floor: dirt	0.84	-	0.85	-	0.85	-	0.84	-	0.84	-
Floor: other	0.16	-	0.15	-	0.15	-	0.16	-	0.16	-
Light: candle	0.20	-	0.62	-	0.77	-	0.83	-	0.59	-
Light: electric	0.06	-	0.05	-	0.06	-	0.05	-	0.06	-
Light: lamps	0.73	-	0.33	-	0.18	-	0.12	-	0.35	-
Roof: grass	0.37	-	0.37	-	0.36	-	0.35	-	0.36	-
Roof: solid	0.63	-	0.63	-	0.64	-	0.65	-	0.64	-
Toilet: flush	0.01	-	0.01	-	0.01	-	0.01	-	0.01	-
Toilet: none	0.09	-	0.06	-	0.06	-	0.04	-	0.06	-
Toilet: pit latrine	0.90	-	0.93	-	0.94	-	0.95	-	0.93	-
Walls: bamboo	0.40	-	0.55	-	0.51	-	0.63	-	0.52	-
Walls: mud	0.39	-	0.26	-	0.35	-	0.21	-	0.31	-
Walls: other	0.10	-	0.04	-	0.02	-	0.04	-	0.05	-
Walls: stone	0.12	-	0.15	-	0.12	-	0.12	-	0.13	-
Water: lake, stream	0.72	-	0.74	-	0.76	-	0.74	-	0.74	-
Water: private tap	0.04	-	0.04	-	0.04	-	0.04	-	0.04	-
Water: public tap	0.10	-	0.08	-	0.08	-	0.06	-	0.08	-
Water: well	0.14	-	0.14	-	0.11	-	0.15	-	0.13	-
Windows: none	0.28	-	0.25	-	0.25	-	0.21	-	0.25	-
Windows: open	0.09	-	0.08	-	0.08	-	0.07	-	0.08	-
Windows: sealed	0.15	-	0.13	-	0.10	-	0.14	-	0.13	-
Windows: shuttered	0.48	-	0.54	-	0.57	-	0.58	-	0.54	-
Bicycles	0.29	0.5	0.33	0.6	0.35	0.6	0.38	0.7	0.34	0.6
Small electronics	0.05	0.3	0.05	0.3	0.05	0.3	0.06	0.3	0.05	0.3
Motorbikes	0.02	0.1	0.02	0.1	0.02	0.2	0.03	0.2	0.02	0.2
Number buildings	1.19	0.5	1.18	0.5	1.18	0.4	1.18	0.5	1.18	0.5
Number rooms	4.72	2.5	4.90	2.4	5.02	2.5	4.96	2.5	4.90	2.5
Radios	0.26	0.5	0.28	0.5	0.31	0.6	0.33	0.6	0.29	0.5
Sewing machines	0.05	0.2	0.06	0.3	0.05	0.2	0.06	0.2	0.06	0.3
Stereos	0.15	0.4	0.16	0.4	0.16	0.4	0.19	0.5	0.17	0.4
Stoves	0.30	0.6	0.03	0.2	0.01	0.1	0.01	0.1	0.09	0.3
Vehicles	0.02	0.2	0.03	0.2	0.02	0.2	0.02	0.2	0.02	0.2

Table 2.11: Factor loadings for the asset index, 1991-1994

	Wave 1	Wave 2	Wave 3	Wave 4	Pooled
Floor: dirt	-0.751	-0.768	-0.764	-0.786	-0.769
Floor: other	0.751	0.768	0.764	0.786	0.769
Light: candle	-0.100	-0.390	-0.351	-0.534	-0.289
Light: electric	0.289	0.273	0.251	0.385	0.288
Light: lamps	-0.066	0.271	0.237	0.357	0.158
Roof: grass	-0.689	-0.664	-0.641	-0.605	-0.665
Roof: solid	0.689	0.664	0.641	0.605	0.665
Toilet: flush	0.287	0.303	0.396	0.326	0.322
Toilet: none	-0.272	-0.216	-0.214	-0.174	-0.224
Toilet: pit latrine	0.157	0.094	0.064	0.032	0.094
Walls: bamboo	-0.286	-0.518	-0.497	-0.604	-0.461
Walls: mud	0.069	0.148	0.179	0.264	0.151
Walls: other	-0.243	-0.161	-0.139	-0.097	-0.171
Walls: stone	0.557	0.632	0.573	0.624	0.599
Water: lake, stream	-0.192	-0.146	-0.225	-0.202	-0.185
Water: private tap	0.478	0.440	0.476	0.422	0.454
Water: public tap	-0.011	-0.040	0.038	0.160	0.026
Water: well	-0.008	-0.042	-0.038	-0.093	-0.049
Windows: none	-0.571	-0.525	-0.508	-0.462	-0.530
Windows: uncovered	-0.191	-0.159	-0.177	-0.185	-0.176
Windows: sealed	0.207	0.211	0.173	0.178	0.199
Windows: shuttered	0.471	0.396	0.438	0.352	0.420
Bicycles	0.414	0.386	0.364	0.324	0.369
Small electronics	0.377	0.459	0.459	0.453	0.428
Motorbikes	0.276	0.290	0.281	0.293	0.281
Number of buildings	0.131	0.175	0.186	0.169	0.162
Number habitable rooms	0.338	0.357	0.398	0.353	0.363
Radios	0.383	0.370	0.326	0.335	0.352
Sewing machines	0.418	0.476	0.425	0.437	0.433
Stereos	0.531	0.528	0.516	0.519	0.519
Stoves	0.474	0.179	0.137	0.191	0.264
Vehicles	0.280	0.362	0.312	0.325	0.317

Notes: unrotated factor loadings from principal factor method with a single factor; factor analysis run separately for each survey wave

Alternative approach to returns estimation

In the main body of the paper we discuss the value of re-estimating the returns equations using categorical variables that represent particular phases of education, rather than the linear educational term. This method allows the marginal value of a year of education to vary across grades; in particular it allows for discontinuities associated with graduation from primary or secondary school. In this section we estimate a standard Mincerian returns function in logs, with dummy variables for different levels of educational attainment, separately for each cross-section:

$$\log E_i = \beta_0 + \beta_1 A_i + \beta_2 A_i^2 + \beta_3 M_i + \sum_{g \in G} [\delta_{mg} M_i D_{gi} + \delta_{fg} (1 - M_i) D_{gi}] + \epsilon_i \quad (2.17)$$

where E_i is earnings in 2010 Tanzania shillings,²⁰ A_i is the age of person i , $M_i = 1$ if person i is male, 0 otherwise, and the vector $[D_{1i} \dots D_{Gi}]$ contains dummy variables corresponding to the educational groups in $G = \{\text{None, Some primary, Primary, Some secondary, Secondary, More than secondary}\}$. We assume $\epsilon \sim N(0, \sigma^2)$, and $E(\epsilon | A, M, D) = 0$. Time subscripts are present but not shown for all components of (2.17), including the parameters. The parameter vector to be estimated via OLS is $(\beta_0, \beta_1, \beta_2, \beta_3, \delta_m, \delta_f)$ where m and f index “male” and “female” respectively, and δ_m and δ_f are 1×6 , with a separate parameter for each educational attainment group.

²⁰2010 prices are used to value consumption bundles in 1991 and 2004.

Table 2.12 gives the results of (2.17). The first column estimates are from the sample of all persons aged 15 or older in the 1991 cross-section. Column 2 was estimated using all persons aged 15 or older in 2004, while column 3 uses only persons aged 20-35 in 2004, because this group was school-aged (age 7-22) in 1991. Columns 4 and 5 are analogous to columns 2 and 3, for the 2010 cross-section rather than the 2004. The estimated coefficients $(a) - (e)$ apply to women; the sum of coefficients $(a) - (e)$ and $(f) - (j)$, respectively, apply to men. The “No education” group is excluded. F -statistics for the overall significance of the education group dummies for men are provided at the bottom of the table. Note that the results in column 3 are included only for completeness. The column 3 sample includes a substantial number of unemployed persons and persons with relatively poor occupational matches. This is because many children who are school-aged in 1991 are still in school, or just out of school and searching for work, in 2004. We therefore restrict our attention to the other 4 columns.

The results in Table 2.12 indicate some clear patterns. The fit of all models is very good for a cross-section in a developing country context, with R^2 ranging from 0.398 in column 1 to 0.514 in column 5. The effect of experience, captured by the coefficients on Age and Age^2 , is essentially stationary. In keeping with most results in developing countries, the expected returns to education are, in general, very high. For women, the expected return to primary education falls from 72% to 54% over the years 1991-2010. The lowest expected return to primary education for women is 40%, in column

Table 2.12: Regression results: log earnings on education and experience

	1991	2004		2010	
	Age 15+ Log E	Age 15+ Log E	Age 20-35 Log E	Age 15+ Log E	Age 26-41 Log E
Some primary (a)	0.347 0.08***	0.216 0.06**	0.314 0.08***	0.276 0.05***	0.197 0.08*
Primary (b)	0.715 0.07***	0.596 0.06***	0.415 0.06***	0.54 0.10**	0.397 0.12*
Some secondary (c)	1.058 0.17***	1.078 0.11***	0.771 0.14***	0.947 0.20**	0.802 0.33*
Secondary (d)	1.051 0.16***	1.137 0.10***	0.955 0.11***	1.13 0.16***	1.144 0.17***
Above secondary (e)	1.761 0.86*	1.185 0.21***	0.989 0.23***	1.618 0.18***	1.555 0.20***
Some primary x Male (f)	-0.174 0.13	-0.473 0.08***	-0.227 0.09*	-0.228 0.04**	-0.081 0.04
Primary x Male (g)	-0.654 0.10***	-0.248 0.07***	-0.189 0.08*	-0.159 0.04**	-0.031 0.03
Some secondary x Male (h)	-0.378 0.27	-0.65 0.11***	-0.451 0.18*	-0.111 0.13	0.02 0.18
Secondary x Male (i)	-0.518 0.25*	-0.385 0.11***	-0.455 0.14**	-0.334 0.10*	-0.381 0.12*
Above secondary x Male (j)	-0.626 0.87	0.051 0.22	0.111 0.27	-0.272 0.13	-0.12 0.08
Age	0.138 0.01***	0.138 0.00***	0.335 0.03***	0.118 0.00***	0.149 0.03**
Age ²	-0.001 0.00***	-0.001 0.00***	-0.005 0.00***	-0.001 0.00***	-0.002 0.00**
Male	1.31 0.10***	1.242 0.06***	1.265 0.07***	1.305 0.03***	1.437 0.02***
R ²	0.398	0.456	0.445	0.466	0.514
N	2466	7001	3613	7161	3586
F: (a) + (f) = 0	3.463	15.606	0.974	1.02	2.621
F: (b) + (g) = 0	0.425	33.684	9.936	24.552	14.189
F: (c) + (h) = 0	13.167	34.138	6.11	55.407	26.541
F: (d) + (i) = 0	11.539	63.749	16.098	11.453	10.152
F: (e) + (j) = 0	51.775	125.819	54.783	21.975	37.926

Note: * sig at 5%, ** sig at 1%, *** sig at 0.1%; standard errors clustered at region level in 2010, village level in 1991 and 2004; constant not shown

5, indicating that for school-aged youth in 1991 the expected return to finishing primary school was lower than it was for their parents.²¹ For women, the same general pattern holds for partial completion of secondary school, though the reduction in expected returns over 1991-2010 is smaller for the overall population than it is in the case of primary completion. Importantly, for women, the expected returns to partial completion of secondary school are substantially greater than the expected returns to primary school across all years.²² This indicates that the expected returns to secondary education are not exclusively due to sheepskin effects or satisfaction of application criteria for high-wage government or private sector jobs: skills attainment in each year of secondary school, and/or ability-based sorting into secondary education, is clearly taking place. The expected returns to secondary education for women increase slightly over the survey period, from 105% in 1991 to about 114% in 2004 and 2010. For women, no clear pattern is discernible for education beyond secondary school, although the expected returns are greater than the expected returns to secondary education in all columns. In general, the marginal expected return to additional schooling at lower levels of female education appears to fall steadily over the period 1991-2010, while the marginal expected return at higher levels of female education increases slightly or remains flat.

²¹This is in keeping both with the increased average levels of educational attainment in Tanzania without concurrent increase in skilled labor employment, and with the structural reforms in the Tanzanian educational system.

²²One-sided *t*-tests reject the null hypothesis of zero difference between coefficients (b) and (c) in all five regressions.

Table 2.13 shows the total effects for men for each educational group. Entries in Table 2.13 are the sums of the appropriate coefficients from Table 2.12. For men, the expected returns to lower levels education do not exhibit the same clear pattern as those for women. In 1991 the expected return to partial primary education, 17%, is greater than the expected return to completed primary education, which is only 6%. This is a puzzling result. One possible interpretation is that men with high-quality land inheritances and/or demonstrated farming aptitude are likely to select out of the final years of primary school. In addition, anecdotal evidence suggests that older generations of Tanzanians grew up with the norm that 4 years of education was sufficient for farmers. Thus, in 1991, the composition of men aged 15+ is a mix of older Tanzanians with the benefits of long years of experience and land acquisition, but with less than primary education on average, and younger adults with less land but higher average education. By 2004 the situation is reversed: partial completion of primary school actually lowers male earnings relative to no education, and the expected return to primary completion is nearly 35%. This is consistent with the interpretation of the 1991 results, since changes in the sample composition from 1991 to 2004 are largely driven by the death of older Tanzanians and the expansion of the sample to include more young and middle-aged adults. In 2010 the expected returns to primary education are even higher, at 38% overall and 36% for the age-group sample, while the expected returns to partial primary school are 4% and 11%, respectively (though the latter two results are not

Table 2.13: Sum of earnings coefficients for men in Table 2.12

	1991	2004		2010	
	Age 15+	Age 15+	Age 20-35	Age 15+	Age 26-41
	Log E	Log E	Log E	Log E	Log E
Some primary x Male	0.173	-0.257	0.087	0.048	0.116
Primary x Male	0.061	0.348	0.226	0.381	0.366
Some secondary x Male	0.680	0.428	0.320	0.836	0.822
Secondary x Male	0.533	0.752	0.500	0.796	0.763
Above secondary x Male	1.135	1.236	1.100	1.346	1.435

Coefficients are the sum of coefficients from previous table. F-stats for significance are in previous table.

significant). The expected return to higher levels of male education follow more intuitive patterns, although we see the same anomaly in 1991 with regard to secondary education as with primary education: partial completion raises expected returns more than full completion. This pattern is reversed in 2004, and reappears in both 2010 columns, but in the latter two columns the difference is insignificant.²³ Importantly, the marginal expected returns for men to secondary and advanced education increase steadily from 1991 to 2010. This is the same pattern that we observed for women.

Once the constants are factored in, the overall trend is that the expected return to education increased slightly over the study period. This is apparent in Figure 2.3, which shows the results from Table 2.12 depicted as the predicted log earnings for men and women, as a function of educational attainment, for each year and sub-group. The vertical intercept is evaluated at age 30 for all figures. The male expected earnings function lies strictly above

²³F-stats from tests of equality of parameters on Some secondary \times Male and Secondary \times Male in columns 4 and 5 are 0.780 and 0.528, respectively.

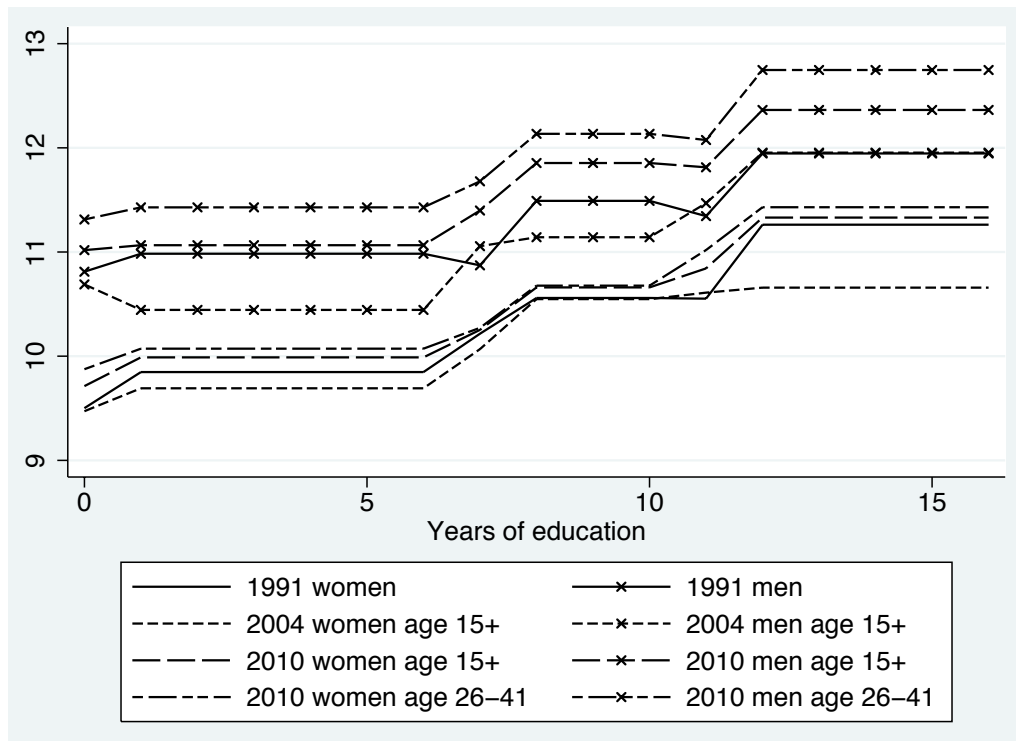


Figure 2.3: Expected log earnings, using the results of Table 2.12

the female expected earnings function in all periods. Earnings in every year is valued using 2010 prices, so the upward trend in both the male and female series reflects real changes in the value of education.

Chapter 3

Using Mobile Phones to Collect Panel Data in Developing Countries

3.1 Introduction

This paper¹ describes the experience of a study entitled Research on Expectations in Agricultural Production (REAP), a survey conducted in rural areas of western Tanzania from July 2009-September 2010. The primary aim of REAP was to gather quantitative data on the evolution and effect of the subjective expectations that farmers hold over uncertain future outcomes,

¹This is the pre-peer reviewed version of an article of the same title which is forthcoming in the *Journal of International Development*, and has been early published in final form at <http://onlinelibrary.wiley.com/doi/10.1002/jid.1771/abstract>.

such as weather, pest intensity and crop yields. Such a project called for high frequency data collection. Instead of embedding enumerators in survey villages for an extended period of time, the REAP team used mobile phones to collect detailed agricultural, economic and demographic data from rural households on a high frequency basis.

The aim of this paper is to describe the mobile phone-based research design and highlight the lessons learned from REAP. Insights presented here can hardly be called best practices, as they are based on the experience of only one project. Nevertheless, it is hoped that this paper will help others avoid some of the challenges that the REAP team has encountered during the planning and execution of a phone-based survey in a remote setting. The paper is organized as follows: in Section 3.2 we describe the project, in Section 3.3 we analyze the methods strengths and weaknesses, and in Section 3.4 we conclude.

3.2 REAP Project Description

During preliminary visits to the study area in 2008 and July 2009, the research team carried phones from each major mobile network in Tanzania, and carefully noted signal availability. We found that one of the network signals was widely available throughout the study area. This did not guarantee network access in every sample village, but it gave us reason for optimism. When the survey began, we were fortunate to find a signal in at least part of

every sample village. The network did not reach some respondents homes, but all households were within a few minutes walk from a signal.

Our sample consisted of 300 cotton farmers in 15 villages. During initial village meetings, we explained the project and provided phone-related training. REAP team members emphasized that the phones were not gifts, but were research tools that would be left in respondents safekeeping. Participants were told that they could use the phones for personal use, and that they could keep the phones once the project was complete. From among the 20 sample farmers in each village, 13 were chosen to participate in the phone survey, for a total of 195 phone survey participants.² To reinforce the notion of random selection, we fully involved the farmers in this stage of selection, by inviting them to draw names from a hat. Prior ownership of a mobile phone did not exclude participants from receiving a project phone.

Phones were distributed on a later day, in the households, after completion of the baseline interview. Respondents also received laminated sheets with the village-specific call schedule and contact numbers for the research team. With 15 villages receiving calls on a Monday-Friday schedule, each village had a calling day once every three weeks.

While none of our sample villages was on the electric grid, some source of power was available everywhere, be it a generator at the school, a house with a solar panel or an individual with a small collection of car batteries. The owners of these power sources operated them as businesses, collecting a

²The seven non-phone households constitute a control group for another study.

small fee to charge a phone. We signed a contract with a charging station in each village, paying for survey participants to receive one free charge during the two days prior to each scheduled call.

From September 2009-July 2010, enumerators called respondents on the prearranged days. We made use of a special block price on within-network calls, paying about \$1.50 per phone for four hours of calls. On most days, one calling block per enumerator was sufficient to complete all interviews. Interview time ranged from 10 minutes to just over an hour, depending on the length of the round-specific questionnaires and the answers given. Average interview time across the 14 rounds of the survey was 27 minutes. Questionnaires included pre-coded, quantitative questions on subjective expectations, labor on- and off-farm, crop sales, livestock sales and purchases, cultivation of cotton and other crops, changes to household composition, health shocks, expenditure on school fees, land holdings, weather, pest intensity, availability of inputs, phone usage, prices, and sources of information. Some of these data were gathered every 3-6 weeks, others less frequently.

Most phone companies will cancel a SIM card if no pre-paid credit is assigned to it for a period of months. Many REAP respondents were unlikely to purchase phone credit on a regular basis, if at all. Both to prevent the cancellation of project SIM cards and to compensate respondents for participating, we transferred 1,000 shillings (about \$0.76) of credit to each phone after each completed interview. The ability to make such transfers is standard in most countries.

We reached an average of eight respondents on the scheduled day. A host of small obstacles prevented interviews from taking place as scheduled, such as illness, family events, network outages, and phone problems. Despite these challenges, virtually all respondents who were not interviewed on schedule were interviewed in the ensuing few days. Village leaders, charging station owners and other participants sometimes assisted us by contacting missing respondents and arranging interviews. Respondents who lost their phones, or whose phones were not working properly, were usually able to participate by borrowing the phone of a friend or neighbor.

A few months after completion of the baseline survey, we re-visited the survey villages and held short meetings with respondents and village leaders. We replaced broken phones, faulty batteries and malfunctioning SIM cards, and topped up our charging station contracts. These visits allowed us to receive additional feedback from respondents, and to demonstrate our commitment to the project.

3.3 Challenges, Solutions and Lessons Learned

In this section I analyze the REAP experience, and speculate more generally on the feasibility of phone surveys in developing countries. I divide the discussion into five subsections: Costs; Infrastructure Issues; Selection and Participation; Data Quality; and Replacement of Materials.

3.3.1 Costs

Relative to a traditional survey, the cost savings from a phone survey are most substantial if the project calls for the collection of panel data over relatively short time horizons. Some field costs cannot be avoided, as researchers must conduct baseline interviews and distribute phones. This involves many of the same budget items as a traditional survey, with the addition of phone-related costs. However, researchers can reduce the time of the initial visit by enumerating sections of the questionnaire that are not time-sensitive at a later time, over the phone. Such an arrangement can reduce field time by days, weeks or even months.

The phones used in REAP cost about 20 each, and each SIM card cost 0.38. The average cost of each of the 2,677 phone survey interviews was \$6.98, including office rental, phone and SIM card purchases, phone charging expenses, air-time, respondent compensation and staff costs. By contrast, the average cost of each of the 195 baseline interviews was approximately \$97, including staff costs, vehicle rental, food and accommodation, printing and other supplies.³ While these estimates are not indicative of the overall cost difference between methods, because they do not assign to the phone survey the costs of the requisite baseline visit, they highlight the key point: once a survey is operational, the marginal cost of gathering additional rounds of data by phone is only a small fraction of what it would be to gather the data

³These cost estimates exclude training, overtime, bonuses, fieldwork permits and some other extra costs, either because such costs were very REAP-specific, or because they applied to both phone and non-phone aspects of the research.

face-to-face.

For the discipline as a whole, phone-based enumeration has the potential to make new types of high frequency data collection feasible in a wide variety of settings, without requiring a substantial increase in funding for field surveys. Individual or household data that may be subject to substantial recall bias in a traditional survey can be gathered more accurately from a high frequency survey. The timing of particular events, such as the employment of inputs or the sale of assets, can be elicited with greater precision. And time-varying data on perceptions and expectations can be gathered in a high frequency panel setting. Such data cannot be reliably gathered with a recall survey. Furthermore, phone surveys are extremely cost effective for research questions that require data at levels of aggregation above the household or individual, such as market price data, quantities available at trading lots or auctions, road or weather conditions.

3.3.2 Infrastructure Issues

Charging the Project Phones

We were fortunate that although none of the REAP villages was connected to the electrical grid, some power source was available in each. Anecdotal evidence suggests that these independent sources of electric power have proliferated alongside mobile phones, in response to the demand for electricity by phone users in rural villages. If true, this bodes well for the feasibility of phone surveys elsewhere. However, a pre-existing source of electricity is

not required for participation in a phone survey. If necessary, researchers can establish charging stations specifically to support the research. Large solar panels cost on the order of \$200-\$500, including installation costs. Alternatively, a number of companies produce small solar chargers for prices as low as \$10, which could be distributed to each participating household.

REAP participants reported a substantial number of faulty batteries. We replaced about 10% of the original batteries during follow-up visits. Some battery problems, such as those caused by irregular voltage from the power source, were unavoidable given the available infrastructure. Other problems, however, were due to a lack of proper training. During the baseline visit, we did not advise participants to turn their phones off when battery power is very low, rather than letting the phones die completely. Nor did we instruct respondents to turn their phones off when outside the network. We also found that some charging station owners took advantage of respondent ignorance by unplugging phones once they display full bars on screen, even though the battery was only 75-80% charged at this point. During follow-up visits we tried to remedy these shortcomings by providing additional training.

Network Access

The limitations of the mobile network may present the most definitive challenge to the feasibility of phone-based data collection. It is impossible to provide a network signal to villages not covered by the existing mobile infrastructure.⁴ Inconsistent mobile network coverage effectively creates a

⁴The only alternative would be to provide respondents with satellite phones, which cost

sampling problem by introducing bias at the village selection stage. This is important for many research questions, since network access is likely to be correlated with other important characteristics, such as distance from major towns, road quality, water supply and average wealth.

Researchers who find that network shortcomings preclude sampling from the original population of interest face tough choices about their project. One possibility is to scrap the phone idea altogether and gather data in the traditional fashion. Another is to draw the sample for the baseline survey from all areas of interest, regardless of network coverage, and then continue the phone survey in those villages with network availability, using characteristics observed during the baseline to construct sample weights. Unfortunately, such weights are only useful if the observables used to construct them are not substantially correlated with network access, and such correlations cannot be measured without first committing to this method of data collection. A third possibility is to establish a calling station as close as possible to a sample village. Unfortunately, this will replace one form of selection bias with another, if the capacity to travel to the calling point is correlated with age, disability, gender, domestic responsibilities, employment status, or other variables of interest.

3.3.3 Selection and Participation

Sampling

\$500-\$1500 each.

We sampled from a list of cotton farmers that we constructed from the official village registry. Village leaders assisted us by removing individuals who had died or moved away, and adding individuals who had moved into the village or formed new households since the most recent registry update. We found that it was best not to mention the phones until after the sample was drawn, to prevent village leaders from tampering with the list in order to increase the likelihood that they or their friends would receive a phone.

For questions related to poverty and household agriculture production, there are few situations in which researchers could justifiably sample from an available list of mobile phone users. SIM cards are inexpensive and widely available, and many phone users own multiple lines. More importantly, phone ownership is highly non-random, and rarely observed among the very poor. To prevent the introduction of substantial sampling bias, REAP was designed with the intention of providing phones to respondents. Although some respondents owned a mobile phone prior to their selection for the study, we did not make use of these phones for REAP, because we did not want to engender ill-will among those who were asked to use their personal phone. Also, although the wealth effect is small, we wanted to endow all participating households with goods of equal liquidity and market value.

In other situations it may be possible to rely on respondents personal phones for enumeration. In Tanzania, phone ownership is nearly ubiquitous among traders, transporters, merchants, university students, government workers and urban formal sector workers. Studies that require sam-

pling from these populations may find that phone distribution is unnecessary. However, attrition rates may be higher under such a design, both because phone endowment appears to engender a deep sense of commitment to the project, and because the opportunity cost of frequent survey participation will be higher among members of wealthier, phone-owning subpopulations.

Attrition and Participation

Potential rates of attrition and periodic non-response⁵ are particularly high when researchers are out of sight for much of the survey period. We anticipated substantial attrition from REAP, due to lack of interest, network problems, or sales of project phones. However, on this point we were pleasantly surprised. Across all rounds of the survey, an average of 191.2 of the 195 respondents were interviewed each round. Missed interviews were due primarily to temporary circumstances, such as severe illness. Only one respondent completely abandoned the survey.

By chance, certain features of the study helped maintain high participation rates. The REAP sampling frame was explicitly restricted to cotton farmers, and we introduced the project as a study of cotton production. Farmers were excited to see interest in their cultivation of the crop they call “white gold” in their language. Also, all of the respondents in the REAP sample lived in relatively small, culturally homogenous villages. Villagers were accustomed to cooperation and neighborliness, and thus were very willing

⁵“Periodic non-response” denotes failure to complete one or more rounds of the survey, while still participating in later rounds.

to help find missing respondents. Such a high degree of cooperation among survey participants was made possible by clustered sampling. It seems very unlikely that response rates would have been so high if respondents had been selected from a higher level of geographic aggregation.

Compensation

Low rates of attrition and non-response were also due to the direct benefits of participation. Many respondents looked forward to the 1,000 Tanzanian shilling (about \$0.76) credit transfer⁶ that they received as compensation for each completed interview.⁷ Some respondents viewed the free battery charge as an additional form of compensation, rather than a practical means of ensuring participation. There was nothing wrong with this perception, however, we were concerned at the outset that some respondents might try to rush through the interview in order to preserve battery life. Fortunately we saw no evidence of such behavior.

Timing

The phone survey did not begin until the 2-month baseline survey was near completion. This introduced a potentially harmful asymmetry into the experiences of respondents, as some waited many weeks for their first phone call, while others waited only a few days. To mitigate the effect of the delay,

⁶These transfers were in the form of air-time credit, which can be used to make calls or send SMS messages, but may also be sold or transferred to other phones.

⁷We learned that it was best to use different phones for calling and for credit transfers. Otherwise an enumerator had to top up her phone with “transfer” credit, exceeding that needed to purchase a daily call bundle. But should she over-run the bundle, the credit intended for transfers was consumed very rapidly at the out-of-bundle rate.

while we were still in the field, enumerators called respondents from the first villages, to greet them and to remind them of their first scheduled calling dates. However, we did not prearrange these calls, and many respondents were unreachable. We learned later that this was largely because their phones did not have any power. This method would have been more successful if we had formally scheduled these calling days, and provided free battery charges. Another way to avoid this problem would be to have a team of phone-based enumerators already in place when the baseline visits begin. Such an arrangement, however, requires enough resources to simultaneously manage data collection in the field and over the phone.

While designing REAP, we also had to decide how often to call respondents. Calling very often over an extended period of time is not only annoying, it is also costly. However, calling too infrequently could raise attrition rates, if respondents lose touch with the project. The optimal lag between calls is clearly related to the research content, the length of the project and the length of each interview. Calling many times a week for two or three weeks is not likely to be such an annoyance as calling many times a week for an entire year. Likewise, interviews that last only a few moments will be tolerated more frequently than those that last close to an hour. The decision to call once every three weeks was made after considering budget constraints, the expected length of each interview, and the nature of the data.

3.3.4 Data Quality

Multiple Languages

It is not uncommon in East Africa for research teams to interview a proportion of respondents in their tribal language, rather than the national language. When necessary, a translator for a face-to-face interview is often selected by the respondent from among his friends and family. In a phone survey, translation can be avoided if an appropriate proportion of the enumerators are fluent in local languages. If a phone enumerator who does not speak a tribal language doubts his ability to communicate effectively with a particular respondent, he or she can transfer the interview to an enumerator who speaks the local language. In practice this was only necessary at the beginning of the REAP phone survey, because during the baseline interview and the first round of REAP calls we identified respondents who were not fluent in Swahili. An enumerator who spoke Kisukuma, the local tribal language, always called these households.

Supervision

The responsibilities of a traditional field survey supervisor generally involve logistics, training and the maintenance of survey quality. In a phone survey, these tasks can usually be accomplished more quickly and at lower expense than in a traditional survey. If interviewers directly enter data into a computer while gathering it over the phone, which is advised for reasons discussed below, questionnaire checking can be automated and performed almost instantly. Supervisors can directly evaluate enumerator performance

by listening to the interviews. In our experience, most community leaders have phones, so the supervisor can remain in contact with local leaders throughout the survey period. All of this can be done from one office, rather than throughout the research areas. The end result is that one supervisor in a phone survey can do the work of many in a traditional survey, without incurring per diem expenses in the field.

Confidentiality and the Interview Environment

Experienced face-to-face enumerators read the body language and facial expressions of the respondent, to see if he is tired, frustrated, confused or intentionally deceptive. Also, traditional interviews are conducted in private, to protect the confidentiality of the data. Unfortunately, phone enumerators cannot observe the respondent during the interview, and they cannot directly ensure confidentiality. This may introduce willful error by a respondent, if the questionnaire content is sensitive. For this reason, researchers studying issues of gender, domestic violence, corruption or other sensitive matters may have difficulty gathering reliable data via phone.

However, to some degree the very nature of a phone interview actually enhances confidentiality. If no one other than the respondent is able to hear the interviewer, and questions require a yes, no, or otherwise innocuous response, respondents can participate in the survey without revealing questionnaire content. The one-sided privacy of a phone conversation is likely not sufficient protection for truly sensitive personal data, but for other topics it may be enough.

Additionally, although a phone interviewer surrenders some control of the interview environment, he surrenders it to the respondent. This can raise response rates and improve data quality, since it allows respondents to easily reschedule interviews. Traditional enumerators often spend substantial time walking to respondents homes. If a respondent is not at home or not in the mood to talk, then a costly re-visit must be scheduled, or the interview must be conducted with an anxious, hurried respondent. These problems are avoidable over the phone.

Data Entry

With regard to data entry, phone surveys seem unambiguously superior to face-to-face paper surveys. Data gathered on paper is transcribed twice: once by the enumerator during the interview, and again by the data entry technician. This creates additional costs, and introduces a delay between data collection and analysis. More importantly, this two-stage transcription of data increases the expected number of errors in the raw data. REAP phone enumerators entered the data directly into a computer during the interview, eliminating the time, expense and potential errors from entry of paper data.⁸

Clarification and Additional Questions

After completion of a traditional field survey, researchers often discover that despite their best efforts, some of the questions were misunderstood by respondents, enumerators or both. Even more frustrating is the realization

⁸The REAP questionnaires were almost exclusively quantitative. Data entry called for input of numeric responses, usually from a menu of pre-coded options, rather than extensive typing of qualitative answers.

that the inclusion of one or two additional questions would have allowed researchers to test unanticipated, yet interesting, hypotheses. Both of these setbacks can be avoided in a phone survey, provided that researchers remain actively engaged with the incoming data. Instantaneous data entry allows identification of potential problems, and interesting new questions, in real time. If REAP enumerators discovered mid-interview that they were unsure of the meaning of a question, or did not know how to handle a particular response, they asked for immediate guidance. Sometimes, enumerators called back respondents to clarify a response. If necessary, a clarification question was inserted into the next round.

3.3.5 Replacement of Materials

It was inevitable that over time, some of the phones and batteries provided by the project would be damaged or lost. Over the nine and half months of the REAP phone survey, eight percent of respondents reported a lost, damaged or malfunctioning phone. These respondents continued to participate in the survey, using the phones of their friends or neighbors. During follow-up visits, research team members replaced most lost or damaged materials.

Replacement of survey materials introduces an element of moral hazard, as respondents are more likely to be careless or to sell the phone and claim that it was lost if they believe it will be replaced. Minimizing the expenses induced by this moral hazard, while still maintaining a spirit of good faith between researchers and respondents, was one of the key challenges of REAP.

To deter sales of the project phones, we told respondent from the outset that we could exchange malfunctioning phones and batteries for new ones, but we could not replace items that were lost. If a respondent lost the project phone but had another phone, we asked him to continue participating in the survey using his personal phone. If a respondent lost the project phone and did not have another phone, we made a determination on a case-by-case basis. We asked about the availability of other phones in the household, and assessed the likelihood of the respondents ongoing participation. If he lived very near to other participants, we usually asked him to continue working with the project using his neighbors phones. However, in a few of these cases we violated our strict policy on replacements, and provided a second phone. We were more likely to replace the phones of those who lived in more isolated areas.

It was clear that some of the lost phones were actually sold. From a research perspective this was not problematic, as long as respondents continued to participate in the survey. It is not clear a priori whether individuals who own another phone are more or less likely to sell the project phone before the survey is complete. The marginal value of a second phone is very low, suggesting that owners of personal phones would be more likely to sell their project phone. However, phone owners are also wealthier in expectation than those who do not own phones, and thus likely to benefit less from a quick sale of the phone for cash. The net effect of these opposing forces is ambiguous.

3.4 Conclusion

On balance, the experience of the REAP study suggests that phone-based enumeration of complex economic surveys in low income countries is not only feasible, but also, under some circumstances, superior to traditional data collection methods. Relative to a traditional survey, the cost savings of a phone survey are substantial, as long as the questions of interest call for high frequency panel data. In addition, the centralized nature of phone-based data collection allows for rapid detection and correction of errors, interactive participation by the primary researchers in real time, and streamlined data entry.

There are situations in which a phone survey is infeasible. Network coverage throughout the study area should be investigated prior to committing to the phone method, so as to prevent the introduction of substantial sampling bias. Elicitation of sensitive data over the phone is unlikely to be successful, as it is impossible for phone enumerators to completely ensure confidentiality. Lastly, its unlikely that the phone survey method will be cost effective for studies that do not require relatively high frequency enumeration of a single set of respondents.

Perhaps the most exciting aspect of mobile phone-based research is the potential it offers for collecting new types of data sets. Current best practices in questionnaire design and data collection methodology are based on the traditional field survey. With the proliferation of mobile telephony comes

the possibility of collecting high frequency panel data at reasonable costs. This should expand the range and number of high frequency panel data sets gathered by development economists, without requiring a large inflow of new research funding.

Chapter 4

Identification of Underlying Beliefs from Subjective Distributions Data

4.1 Introduction

Subjective probabilities over uncertain future outcomes occupy a prominent place in the standard von Neumann-Morgenstern (1947) model of expected utility, and in other theories of choice under uncertainty. Nevertheless, in the second half of the 20th century, theoretical and empirical economic research concentrated on modeling and estimating expectations in observed choice data, rather than directly gathering subjective probabilities from economic agents. Manski (2004) speculates on possible reasons for the taboo

among economists against the collection and analysis of subjective expectations data. Whatever the reasons may have been, over the last decade the taboo has lifted, and significant efforts have been made to directly measure agents' expectations rather than impose them on choice data by assuming rational, adaptive, or other expectations processes.

A number of widely used data sets from the US, such as the Health and Retirement Survey, the National Longitudinal Survey of Youth, the Michigan Survey of Consumers and the Survey of Economic Expectations, include questions about subjective point expectations or distributions. Questionnaire modules to elicit subjective probability distributions in these surveys typically gather between 2 and 10 points on the cumulative distribution function or probability density function of person j 's unobserved belief about the distribution of random variable z . In the Survey of Economic Expectations (see Dominitz 1998, Dominitz and Manski 1997), this was accomplished by asking respondents the likelihood of their income in some future period falling below fixed values $\{z_1, \dots, z_N\}$, with $z_1 < z_2 < \dots < z_N$.

More relevant for the current paper is the rapid growth of the development economics literature on the measurement and analysis of subjective distributions data.¹ In data sets gathered in low income countries, visual aids are usually used to elicit distributional data from populations with lower average levels of education. The method that has quickly become standard is to ask respondents to allocate a fixed number of counters or other counters

¹See Delavande *et al* (2010) for a review.

to boxes that represent the intervals of a histogram. The proportion of counters allocated to each “bin” represents the density in that interval. The same method is often used when the boxes represent qualitative outcomes, such as the values of a Likert scale (“Very bad”, “Bad”, “Ok”, “Good”, “Very good”).

Subjective distributions data gathered with either of these methods are different from other survey data in one critical respect: even perfectly formed beliefs about the distribution of an unknown variable cannot be communicated from the respondent to the researcher. Given the limitations on survey time, researchers can only collect a finite number of points on the PDF or CDF of the stochastic variable. The questions used to collect these data are in essence choice problems that the researcher poses to the respondent. In a strict sense, all survey data can be described as the outcome of a choice problem. The respondent must choose an answer to report, and he must decide how much effort to exert recalling bygone events. In a lab or field experiment, he must make choices that we hope reflect the same underlying preferences that he uses to make choices in authentic markets. In these scenarios, however, the respondent is generally able to provide a full and complete answer to the question if he so chooses, because he is not attempting to map infinite-dimensional information into a finite number of answers.

When providing subjective probability distributions, however, that is precisely what the respondent is doing. There is simply no way for the respon-

dent to fully communicate a non-degenerate belief to the researcher.² Therefore, in order to analyze the information content of subjective probability distributions survey data, we first have to understand the choice problem solved by the respondent. Our understanding of this problem, and its effect on the identification of the respondent's underlying belief, has implications both for survey design *ex ante* and data analysis *ex post*.

In this paper, we focus on the bin-and-counter choice problem most commonly used to gather subjective distributional data in low income countries.³ Data gathered using this method partially identifies the underlying subjective distribution, the CDF of which we denote by $F_j(z)$, and the PDF by $f_j(z)$. Our main goal is to characterize this partial identification problem. To date, researchers have employed a wide range of different tactics to recover moments and/or fit continuous approximations to these “binned” data. None of these tactics is based on rigorous consideration of the underlying partial identification problem. This has serious consequences for the findings of any research project that employs these data, because the higher central moments of the estimated subjective distribution - “higher” being anything greater than the first - are very sensitive to distributional assumptions. Table

²There is substantial work in psychology and marketing on the cognitive aspects of the survey response process, which have additional implications for the interpretation of survey data. See McFadden *et al* (2005) for a discussion. We momentarily ignore these additional complicating factors in order to emphasize that even in a world with fully competent, self-aware and honest respondents, non-degenerate subjective probability distributions data cannot be communicated in their entirety from one person to another.

³As we will see, data gathered in this fashion have lower information content than data gathered in the manner described above for the Survey of Economic Expectations. The latter will be a limiting case of the bin-and-counter data.

Table 4.1: Subjective Distributions in Some Recent Papers

Reference	Random variable	N	k	Method of constructing bins	Distributional assumptions
Hill (2010)	Coffee prices	3	20	Fixed by researcher	Density concentrated at midpoint of bin
Delavande et al (2010b)	Fishermen's daily catch	10	10,20	Fixed by researcher	Stepwise uniform
Cole and Hunt (2010)	Commodity prices	5	20	Elicit $E(p)$, construct quartiles around $E(p)$	Stepwise uniform
Attanasio and Kaufmann (2009)	Income	2	100?	Elicit max, elicit min, divide in half	Stepwise uniform, triangular, bi-triangular
Gine et al (2008)	Monsoon timing	11	10	Fixed by researcher	Each stone is a random draw from beliefs distribution
McKenzie et al (2007)	Income after migration	4	NA	Elicit max, elicit min, divide into quartiles	CDF method, log normal wages

4.1 lists the random variables of interest and the distributional assumptions employed in a number of published and working papers that use subjective distributions data. The lack of standard practice is evident.

There are no asymptotics in this paper, because the data in question are generated by a single respondent (or, more accurately, many respondents, but each one considered in isolation) assigning a small number of counters (usually 10-20) to a small number of bins (anywhere from 2 to 10). We are not concerned with any population-level phenomena, but with the identification of a single respondent's belief about the distribution of an uncertain outcome. We focus on the individual-specific distribution because data of this type are most often used to better understand choice behavior by exploiting observed heterogeneity in subjective beliefs. This is in contrast to the literature that uses probabilistic information gathered from experts to make inference about

an unobservable population distribution (see Kiefer 2010).

Despite the apparent similarity to interval data, subjective distributions data are inter-dependent, and thus do not satisfy most of the properties of standard observational data. With standard interval data, each data point represents a single independent draw from a population, and the appropriate interval for any unobserved value is invariant to changes in the number of intervals or boundaries of the intervals to which it does not belong. However, the interval data generated by a single person allocating counters to bins is not invariant to changes in the number of bins, the number of counters, or the boundaries of the bins. When considering where to place a counter, the respondent compares the bins to each other and determines, in essence, the *relative* likelihood of the unknown outcome taking on a value in each interval. For this reason, recent developments in the partial identification of interval data are not directly applicable to the current problem.⁴ In the next section, after formalizing the respondent's allocation problem, we consider an example that makes this point more clear.

This paper makes a number of contributions to the analysis of subjective distributions data. We first provide evidence that a number of reasonable heuristics that a respondent might follow when allocating the counters are all equivalent to the minimization of absolute loss between the allocation and the unobserved $f_j(z)$. We then provide bounds on the density in any subset of the

⁴The results in Stoye (2010), however, apply to some types of subjective distributions data. We will elaborate below.

bins, bounds on the subjective CDF, and a complete characterization of the joint identification region for the vector of unobserved densities, as a function of the numbers of counters and bins. The boundaries of this identified region are sharp in the sense that they exhaust all of the information provided by the respondent. I define a non-parametric estimator that can be used to bound unobserved information signals when subjective distributions are gathered in a panel. The identification region for the measure vector is then used to generate joint bounds on the moments of $f_j(z)$. Lastly, we provide Monte Carlo evidence for the appropriate way to fit a continuous approximation to a respondent's allocation of the counters, and simulation results of the effect of different numbers of counters and bins on the size of the mean-variance identification region.⁵

The paper proceeds as follows. In the following section we consider the respondent's choice problem, and argue that a minimization of absolute loss allocation rule is consistent with numerous allocation heuristics that the respondent might use. In section 4.3 we derive bounds on the density in subsets of the bins, bounds on the CDF and a characterization of the joint identification region. Section 4.4 describes a numerical method for joint identification of the expectation and variance of $f_j(z)$, conditional on the response. In section 4.5 we provide simulation evidence both for the choice of smoothing technique *ex post* and the selection of the number of bins and counters *ex*

⁵Estimation of a subjective distribution that follows a known parametric form may be desirable for use in a structural model, or for mixing two subjective distributions (such as price expectations and quantity expectations).

ante. Section 4.6 concludes.

4.2 The Respondent's Choice Problem

4.2.1 Preliminaries

Suppose that a respondent is presented with N bins, each representing an interval on the support of the distribution of unknown outcome $z \in \mathbb{R}$. The unknown z may be continuous or discrete. Bin i is defined as $[\underline{d}_i, \bar{d}_i]$, $i = 1 \dots N$, with $\bar{d}_i = \underline{d}_{i+1}$ for $i = 1 \dots N - 1$. Extensions to situations in which $\bar{d}_i \neq \underline{d}_{i+1}$ are straightforward, though less common empirically. It does not matter whether the bins are of equal size. Call the vector of lower bounds $\underline{d} \in \mathbb{R}^N$, and the vector of upper bounds $\bar{d} \in \mathbb{R}^N$.

From here onwards we suppress the j subscripts denoting a particular respondent. A respondent's belief about the distribution of z is denoted by density function $f(z)$, which satisfies the usual assumptions. The support of $f(z)$ is bounded by $[\underline{a}, \bar{a}]$, so that $\int_{\underline{a}}^{\bar{a}} f(z) dz = 1$. I make the additional assumption that $\underline{d}_1 \leq \underline{a} < \bar{a} \leq \bar{d}_N$, i.e. the entire positive mass of $f(z)$ lies within the interval covered by the bins. This may not be the case in practice, if the respondent places positive probability on values outside the range anticipated by the researcher. However, if z has natural bounds (such as a non-negativity constraint), or if the lowermost and uppermost edges of the visual aid are left open-ended, the researcher can check results for various choices of \underline{d}_1 and \bar{d}_N to ensure that results are not overly sensitive to change

in the assumed boundaries of z .⁶

The respondent is given $k > N$ counters and asked to allocate them among the N bins, so as to represent his belief $f(z)$ as closely as possible. One can think of the value $\frac{1}{k} \in (0, 1)$ as the “raw value” of a counter, the proportion of the total probability represented by each counter. However, we will see below that in general, not all counters represent this exact proportion of the total probability.

Define an *allocation* $x = (x_1, \dots, x_N)$ to be the respondent’s division of the k counters among the N bins, such that:

1. $x_i \in \mathbb{Z}_+, \quad i = 1 \dots N$
2. $\sum_{i=1}^N x_i = k$

In order to coherently make such an allocation, the respondent must parse his belief $f(z)$ into an N -vector of probabilities $p = (p_1, \dots, p_N)$, satisfying:

1. $p_i = \int_{\underline{d}_i}^{\bar{d}_i} f(z) dz$
2. $p_i \in [0, 1] \quad i = 1 \dots N$
3. $\sum_{i=1}^N p_i = 1$

where 2. and 3. follow from the definition of $f(z)$. Each p_i represents the respondent’s belief about the likelihood of z taking on a value in bin

⁶The estimated higher moments of $f(z)$ can be very sensitive to assumptions about the support, especially if the respondent places counters in the first or last bin. The first best solution is to develop a visual aid with uppermost and lowermost intervals that are very unlikely to receive a counter.

i. Note that we can split each element of p into two components: $p_i = (\text{floor}(kp_i) \cdot (\frac{1}{k})) + r_i \equiv q_i + r_i$, where the function $\text{floor} : \mathbb{R}_+ \rightarrow \mathbb{Z}_+$ is defined as $\text{floor}(r) = \max\{z \in \mathbb{Z}_+ | z \leq r\}$. The first component q_i is a multiple of the raw value of a single counter. The second component $r_i \in [0, \frac{1}{k})$ is the “residual” or the “remainder”, the density in bin i over and above that which can be accounted for by an integer number of counters. Denote $r \in [0, \frac{1}{k})^N$ as the *residual vector*.

Hereinafter we refer to the probability vector p as a *measure*. We assume that the respondent is able to conceptualize the measure that is consistent with his belief $f(z)$, and that he does not make any errors in allocating the counters. In so doing, we assume that the respondent does not extract any information from the visual aid when making his allocation. In practice, a respondent may adjust his subjective expectations on the spot if he is presented with a visual aid that suggests, implicitly, that z has a support very different from that of his prior $f(z)$. Such a response would be similar to the “bracketing effects” discussed in the psychology literature (Schwartz *et al* 1998). While it is plausible that a respondent might be influenced by such an effect when making his allocation, this paper does not take up this possibility formally.

Lastly, define a *location* $w \in \mathbb{R}^N$ to be a vector of values satisfying $w_i \in [\underline{d}_i, \bar{d}_i]$ for $i = 1 \dots N$. A location is an N -vector with a single element in each bin. Different locations correspond to different discrete distributions for which the measure is “stacked” at different points in $[\underline{d}_i, \bar{d}_i]$, $i \in \{1, \dots, N\}$.

There is no reason to think that beliefs are, in general, discrete. However, it is easy to see that the bounding values in the sections to follow - bounds on bin probabilities, on the cumulative distribution function of $f(z)$, and on the moments of $f(z)$ - are associated with discrete distributions for which some or all elements of w are located at the boundaries of their respective bins. We therefore restrict attention in much of the paper to discrete distributions, characterized by a measure p and a location w .

As is clear from the definition of p , there is a single measure that represents the respondent's belief $f(z)$. In non-boundary cases, there is also a single allocation x that represent p . For finite k , neither representation is 1:1. Therefore, there are infinite beliefs f that correspond to any observed allocation x . The researcher's task is to learn as much as possible about $f(z)$ from the observation of x , by bounding the set of measures p that are consistent⁷ with an observed allocation.

4.2.2 Four Allocation Heuristics

In order to make progress on this problem, we first have to understand the decision rule that the respondent uses to allocate the counters to the bins. With no assumptions about the allocation process, none of the order, spread or location statistics of f is identified. Absent the assumption made above, that $f(z)$ is entirely contained by the range of the bins, none of these statistics

⁷There are no asymptotics in this paper, so I use the term "consistent" in the non-statistical sense.

is even bounded.

Consider the following four allocation heuristics, all of which have been discussed in the literature or suggested by other researchers:

A1. Iterative Allocation

In order to track the iteration, rename vector p as p^0 . The respondent places a counter in bin j such that $p_j^0 = \max_{i=1,\dots,N}\{p_i^0\}$. If the maximum is not unique, the respondent randomly chooses one of the maximum-probability bins. He then forms the vector p^1 , with $p_i^1 = p_i^0$ for $i \neq j$, $p_j^1 = \max\{p_j^0 - \frac{1}{k}, 0\}$, and repeats the process by placing the second counter in the new maximum-probability bin. He iterates until all k counters have been placed.

A2. Initial and Residual Allocation

The respondent places the counters in the two stages. He allocates $\text{floor}(kp_i) = kq_i$ counters to each bin in the “initial stage”. There will be $k_r \in \{0, \dots, N\}$ counters left over. Unless $kp_i \in \mathbb{Z}^+ \forall i$, it will be the case that $k_r > 0$. In the second “residual stage”, the respondent allocates the remaining counters to the k_r bins with the highest valued residuals r_i . As in [A1.], he randomizes in the event of a tie.

A3. Rounding

Before allocating counters, the respondent rounds each p_i to the nearest $\frac{1}{k}$. This gives the *rounded measure* \bar{p} , for which $\bar{p}_i = q_i$ if $r_i < \frac{1}{2k}$, and $\bar{p}_i = q_i + \frac{1}{k}$ otherwise. For example, if $k = 20$, the respondent

rounds the measure in each bin to the nearest 0.05. If $\sum_{i=1}^N \bar{p}_i = 1$, which will not always be the case, the respondent places $k\bar{p}_i$ counters in each bin i . If $\sum_{i=1}^N \bar{p}_i < 1$, each bin receives $k\bar{p}_i$ counters, and the remaining $k(1 - \sum_{i=1}^N \bar{p}_i)$ counters are allocated to those bins j for which $\bar{p}_j = q_j$ and r_j is greatest (that is, some of the “rounded down” bins are given an extra counter, as if they had been rounded up). Similarly, if $\sum_{i=1}^N \bar{p}_i > 1$, each bin receives kq_i counters, and the remaining $k(1 - \sum_{i=1}^N q_i)$ counters are allocated to those bins j for which r_j is greatest (some of the “rounded up” bins are rounded back down).

A4. Uniform Opening, with Redistribution

The respondent begins in stage $s = 0$ with a uniform allocation, by placing $x_i^0 = \frac{k}{N}$ counters in each bin i . Suppose for simplicity that $\frac{k}{N} \in \mathbb{Z}_{++}$, so that this allocation is feasible.⁸ The respondent considers the absolute loss in stage s , given by $L^s = \sum_{i=1}^N |\frac{x_i^s}{k} - p_i|$. He classifies the bins into two sets, the set of “over-weighted” bins $I_l^s = \{i \in 1, \dots, N \mid p_i < \frac{x_i^s}{k}\}$, and the set of “under-weighted” bins $I_u^s = \{i \in 1, \dots, N \mid p_i \geq \frac{x_i^s}{k}\}$. In each of these sets define the most over-weighted (under-weighted) bin as, respectively, $x_i^{ls} = \{x_j^s \mid j \in I_l^s \text{ and } \frac{x_j^s}{k} - p_j = \max_{c \in I_l^s} \{\frac{x_c^s}{k} - p_c\}\}$, and $x_i^{us} = \{x_j^s \mid j \in I_u^s \text{ and } p_j - \frac{x_j^s}{k} = \max_{c \in I_u^s} \{p_c - \frac{x_c^s}{k}\}\}$. In stage $s = 0$, x_i^{l0} is associated with the bin that has the minimum measure p_i across all bins, and x_i^{u0} is associated with the bin that has the maximum mea-

⁸All results go through if this assumption is relaxed, but by invoking it we shorten an already arduous and repetitive proof in the Appendix.

sure, because these are the bins with measure on either side of initial measure $\frac{k}{N}$ that contribute most to the loss under a uniform allocation. To move to stage $s = 1$, the respondent re-allocates one counter from x_i^{l0} to x_i^{u0} , so that $x_i^{l1} = x_i^{l0} - 1$ and $x_i^{u1} = x_i^{u0} + 1$, and re-calculates the loss. If $L^1 < L^0$, he repeats the process, starting at $s = 1$. Otherwise, he returns the counter and considers the allocation complete.

Consider an example, in which the respondent employs heuristic [A2.]. Suppose that $N = 3$, $k = 8$, and $p = (0.3, 0.5, 0.2)$. The raw value of a counter is $\frac{1}{8} = 0.125$. In the initial stage, the respondent makes the initial allocation $(2, 4, 1)$, to account for the densities $(0.25, 0.5, 0.125)$ that are multiples of 0.125. One counter remains un-allocated. The residual vector is $(0.05, 0, 0.075)$, so the final counter is placed in the third bin, resulting in final allocation $x = (2, 4, 2)$.

In fact, facing the same situation, a respondent using any of the other three allocation heuristics defined above would provide the same allocation, $x = (2, 4, 2)$. Proposition 1 formalizes the equivalency between these allocation heuristics.

Proposition 1. *For any measure p corresponding to underlying beliefs $f(z)$, heuristics [A1.]-[A4.] induce the same allocation x , up to the randomization that occurs if more than one bin satisfies the conditions for receipt of the marginal counter. Furthermore, a respondent employing any one of these heuristics allocates the counters so as to minimize absolute loss, given by $L = \sum_{i=1}^N |\frac{x_i}{k} - p_i|$, subject to the restriction that $\sum_{i=1}^N x_i = k$.*

A proof of Proposition 1 is given in the Appendix.

Proposition 1 provides a compelling rationale for treating the data as if the respondent takes his own beliefs p as given, and chooses the allocation x so as to minimize absolute loss $L = \sum_{i=1}^N |\frac{x_i}{k} - p_i|$, conditional on the requirement that all counters be allocated. As we have just seen, this allocation rule is consistent with a wide range of heuristics. Although the logic underlying this paper could be used to study other allocative processes,⁹ the formal results depend critically on the assumed allocation rule. The allocation rule defines the relationship between the unobserved measure p and the observed allocation x , and so provides the channel through which we can analyze the identification of $f(z)$.

The condition in Proposition 1 that $\sum_{i=1}^N x_i = k$, which requires the respondent to allocate all of the counters, will in some cases preclude the unconditional minimization of absolute loss. For example, suppose that $N = 3$, $k = 10$, and $p = (0.92, 0.04, 0.04)$. After placing nine counters in the first bin, one counter remains, and absolute loss prior to placing the final counter is $|0.02| + |0.04| + |0.04| = 0.1$. However, the minimum value of absolute loss after the final counter is allocated is 0.12.¹⁰ Thus, respondents using the above allocation heuristics are minimizing absolute loss subject to

⁹One possibility is that respondents penalize the density in high probability bins so as to place a counter in bins with low, but non-zero probability. For example, after placing 9 of 10 counters in bin i for which $p_i = 0.96$, a respondent may feel compelled to place the final counter in bin j with $p_j = 0.04$, to indicate that he does place some positive probability in a bin other than i . This is a sort of reverse probability-weighting. Another possibility relates to the bracketing effects already discussed. We do not study either of these possibilities formally, yet both are conceivable deviations from the fully rational minimization of absolute loss rule.

¹⁰The final counter can be placed in either bin 2 or bin 3.

the requirement that all counters be placed. We maintain this restriction because, to date, those gathering subjective distributions data of this type have always required respondents to allocate all of the counters.

In the discussion to follow I will primarily use the language of heuristic [A2.], that of “initial stage” and “residual stage” allocations. I use this terminology because it emphasizes a key component of the identification problem. Recall that we can write $p_i = q_i + r_i$, where q_i is divisible by $(\frac{1}{k})$. Unless $r_i = 0$, all counters but the last one placed in any given bin are placed to account for the measure q_i . Each of these counters represents the exact density $\frac{1}{k}$, the raw value of a counter. The final counter in any bin i , however, is placed because the residual probability in bin i is sufficiently great, relative to the residual probabilities in the other bins, to merit placement of another counter. Most of the bounds derived below are determined by considering the extreme values of residual vector r that are consistent with observed allocation x . Thus, it is the residual stage allocation that receives most of our attention.

4.2.3 Implications of Respondents’ Decision Rule

Before moving on, we consider the implications of the allocation rule for the relationship between p and x . Let $X_0 = \{p_i | x_i = 0\}$ indicate the set of bin-measures corresponding to the “empty bins”. Likewise, let $X_1 = \{p_j | x_j > 0\}$ be the set of measures corresponding to the non-empty bins. We make this distinction because the measure in any empty bin is naturally bounded below

by 0, whereas the measure in any non-empty bin is not. We will use N_0 to denote the number of empty bins, and N_1 to denote the number of non-empty bins, so that $N_0 + N_1 = N$. Let $l_i^s = |\frac{x_i}{k} - p_i|$ be the contribution to the absolute loss from bin i .

Proposition 2 formalizes some of the relationships between the measure in empty and non-empty bins, and between p and x , that are implied by minimization of absolute loss.

Proposition 2. *Minimization of absolute loss during allocation of the counters implies the following:*

$$x_i \in \{kq_i, kq_i + 1\} \subset \mathbb{Z}_+ \quad \text{for all bins } i \quad (4.1)$$

$$l_i = \begin{cases} r_i & \text{if } x_i = kq_i \\ \frac{1}{k} - r_i & \text{if } x_i = kq_i + 1 \end{cases} \quad (4.2)$$

$$p_i \leq p_j \quad \text{for any } p_i \in X_0, p_j \in X_1 \quad (4.3)$$

$$p_j > \frac{x_j - 1}{k} \quad \text{for any } p_j \in X_1 \quad (4.4)$$

$$p_j - p_l \leq \frac{x_j - x_l}{k} + \frac{1}{k} \quad \text{for any } p_j, p_l \in X_1 \quad (4.5)$$

Statement (4.1) indicates that each bin i receives either $\text{floor}(kp_i)$ or $\text{floor}(kp_i) + 1$ counters. Statement (4.2), which asserts that the value of the error from each bin is a function of the residual measure in that bin, follows immediately. Statement (4.3) is trivial. Statement (4.4) is equally trivial, but it is provided to emphasize the fact that all but the final counter placed in non-empty bin j accounts for measure $\frac{x_j - 1}{k}$. Statement 4.5 is critical in the analysis below, as it bounds the difference between the measure in any two non-empty bins.

Figure 4.1 gives a stylized example that emphasizes the inter-dependence discussed in the introduction. Suppose that two respondents have the same underlying beliefs $f(z)$. One is presented with the visual aid shown on the left, for which $N = 2$. The other is presented with the visual aid on the right, for which $N = 3$. Both are given $k = 4$ counters to allocate. Note that the bounds of bin 1 are the same in both cases. Suppose furthermore that the boundaries of the bins are placed such that beliefs $f(z)$ correspond to measure $p = (0.34, 0.66)$ in panel A, and to $p = (0.34, 0.33, 0.33)$ in panel B. Under the absolute loss rule, the first respondent provides allocation $x = (1, 3)$, while the second provides allocation $x = (2, 1, 1)$. A naive read of the data might lead one to conclude that respondent 2 believes $Pr(z < \bar{d}_1)$ is twice as great as does respondent 1, even though the underlying beliefs, the number of counters and the values \underline{d}_1 and \bar{d}_1 are the same in both cases. This example highlights the importance of thinking hard about the interdependence between N , k and the allocation process, before making inferences about $f(z)$.

Before moving on, it is useful to lay to rest another concept borrowed from experience with observational data. Intuition may suggest that symmetric beliefs $f(z)$ induce symmetric allocations x , and that, consequently, the observation of a symmetric allocation x is evidence of symmetric underlying beliefs f .¹¹ However, from an identification perspective, there is no positive relationship between symmetry of allocations and symmetry of beliefs. If the

¹¹We use the term “symmetry” exactly as one would expect. For example, $(2, 4, 2)$ and $(4, 0, 4)$ are symmetric allocations, while $(1, 1, 6)$ is not.

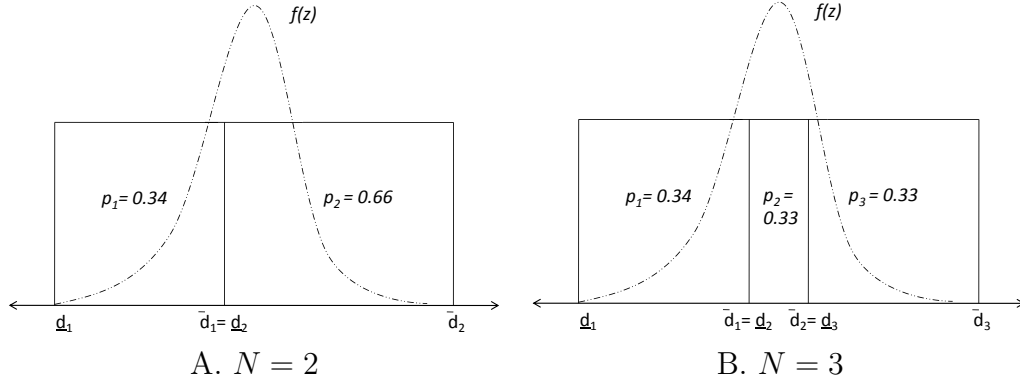


Figure 4.1: Measure parsing with identical beliefs, for $N = 2$ and $N = 3$

boundaries of the bins are “off-center” with respect to the median of symmetric $f(z)$, beliefs may be represented with a highly skewed allocation. In the sections to follow I will repeatedly use an example in which underlying beliefs are symmetric, but the allocation is not. Asymmetric allocations do in some cases *exclude* the possibility of symmetric beliefs; but asymmetry of beliefs can never be excluded.

4.3 Bounds on Bin Probabilities

In this section we identify bounds on the elements of $p \in [0, 1]^N$ that are consistent with a given allocation $x \in \mathbb{Z}_+^N$. These bounds are functions of N and k . As intuition suggests, the bounds narrow as the number of bins and/or the number of counters increases. We first consider the bounds on the measure in any single bin. From there it is a simple extension to provide joint bounds on any subset of the bins, which allows us to bound the CDF of the

underlying $f(z)$. Lastly, we fully characterize the identified set $\mathbb{P} \in [0, 1]^N$.

4.3.1 Single bin bounds

Statement (4.4) in Proposition 2 highlights the point that when considering bounds on p , we need only consider the residual stage allocation. We can therefore divide the full measure $P = 1$ into initial measure P_I and residual measure P_R , such that $P_I + P_R = 1$. These are defined as follows:

$$P_I = \sum_{j \in X_1} \frac{x_j - 1}{k} \quad (4.6)$$

$$P_R = 1 - P_I = \frac{N_1}{k} \quad (4.7)$$

The initial measure P_I is that which is accounted for by all but the final counters placed in the bins. There is nothing uncertain about such measure, because it would not have been optimal to place an additional counter in bin i if $p_i \leq \frac{x_i - 1}{k}$. The residual measure P_R is the object of interest. In order to distinguish between empty and non-empty bins, let $A = \{p_i | i \in X_0\}$ be the set of contributions to P_R from the N_0 empty bins, and let $B = \{p_j - (\frac{x_j - 1}{k}) | j \in X_1\}$ be the set of contributions to P_R from the N_1 non-empty bins. These have typical elements a_i and b_i , respectively, which we refer to as *residual measures*.

Note that P_I and P_R are defined from the perspective of the researcher, not the respondent. Thus the identified region will include measures p for

which the final counter placed in any bin was an initial placement, rather than a residual placement.¹² However, measures for which the *second-to-last* counter placed in any bin was placed to account for the residual r_i are always excludable from the identified region.

As an example, suppose that $N = 5$, $k = 10$, and the bins are the non-overlapping intervals of width 200 spanning $[0, 1000] \subset \mathbb{R}$. Suppose the respondent provides the allocation $x = (0, 1, 5, 4, 0)$ (this will be the “standard example” that we follow below). We know with certainty that $p \geq \tilde{p} = (0, 0, 0.4, 0.3, 0)$. Then the initial measure is $P_I = 0.7$, and the residual measure is $P_R = 0.3$. In considering the identification problem, our focus is on the placement of the final 3 counters.

Bounds on empty bin measures

It should be clear without proof that the minimum residual measure in any empty bin i is $a_i = 0$. Consider the researcher’s problem to identify the maximum value of the residual measure a_i in an empty bin i . Note that maximization (minimization) of a_i is tantamount to maximization (minimization) of p_i , for $i \in X_0$. Given allocation x , the researcher’s maximization problem

¹²Clearly, such bins did not receive a counter in the residual allocation.

can be written as follows:

$$\max_{(a_1, \dots, a_{N_0}, b_1, \dots, b_{N_1}) \in [0,1]^N} a_i \quad \text{s.t.} \quad P_r = \sum_{m=1}^{N_0} a_m + \sum_{j=1}^{N_1} b_j \quad (4.8)$$

$$b_j - b_l \leq \frac{1}{k} \quad \forall j, l \in X_1 \quad (4.9)$$

$$a_m \leq b_j \quad \forall m \in X_0, j \in X_1 \quad (4.10)$$

$$a_m, b_j \geq 0 \quad \forall m \in X_0, j \in X_1 \quad (4.11)$$

Constraints (4.9)-(4.11) follow directly from Proposition 2. Note that x only enters the problem indirectly, via N_0 and N_1 . The objective function in (4.8) is linear and the bounds defined by the constraints constitute a closed, compact domain $D \subset [0,1]^N$ for choice vector (a, b) . Therefore, by the Weierstrass theorem, the objective function $g_a(a, b) = a_i$ achieves a maximum on D .

The solution to the problem in (4.8)-(4.11) is derived in the Appendix. The residual vector at the optimum is given by:

$$\begin{aligned} a_i^* &= \frac{N_1}{k(N_1+1)} \\ b_j^* &= \frac{N_1}{k(N_1+1)} \quad \forall j \in X_1 \\ a_l^* &= 0 \quad \forall l \in X_0 \setminus i \end{aligned} \quad (4.12)$$

which leads immediately to Proposition 3:

Proposition 3. *Given any allocation $x \in \mathbb{Z}_+^N$ and the ensuing values $N_0, N_1 \in \mathbb{Z}_+$, the measure p_i in any empty bin i satisfies $p_i \in [0, \frac{N_1}{k(N_1+1)}]$.*

The intuition behind this result is clear. Recall that $P_R = \frac{N_1}{k}$, so the upper bound in Proposition 3 can be written as $\frac{P_R}{N_1+1}$. Consider the standard example, with $x = (0, 1, 5, 4, 0)$. Recall that we are only interested in the allocation of the final counter to each of bins 2, 3 and 4, and the $P_R = 0.3$ residual probability that they represent. Suppose we want to know the maximum value of p_1 that is consistent with the observed x . The maximum value of empty bin 1 is that for which the other empty bin has measure 0, i.e., $p_5 = 0$, and the final counter placed in bins 2, 3 and 4 could have been placed in bin 1, but was not due to sheer randomization. This corresponds to the beliefs for which the residual probability $P_R = 0.3$ is distributed uniformly across the $N_1 + 1$ bins 1, 2 3 and 4, i.e., $p_1 = \frac{0.3}{4} = 0.075$. Therefore, the unique vector of beliefs that includes the maximum identified value of p_1 , and induces allocation $x = (0, 1, 5, 4, 0)$, is $p = (0.075, 0.075, 0.475, 0.375, 0)$.

Bounds on non-empty bin measures

We turn now to the bounds on the measures in the non-empty bins. The example just provided for empty bins provides helpful intuition. The lower bound on the measure in any non-empty bin i is the measure for which the final counter *could have been placed somewhere else*, but bin i was randomly chosen from the set of bins that merited the marginal counter. Similarly, the upper bound on p_i for non-empty bin i is the measure which was randomly selected to *not* receive an additional counter, even though residual probability r_i made it a candidate to receive one more counter.

Using the notation from the previous section, the researcher's maximization problem for the non-empty residual measure b_i is similar to that for a_i :

$$\max_{(a_1, \dots, a_{N_0}, b_1, \dots, b_{N_1}) \in [0,1]^N} b_i \quad \text{s.t.} \quad P_r = \sum_{m=1}^{N_0} a_m + \sum_{j=1}^{N_1} b_j \quad (4.13)$$

$$b_j - b_l \leq \frac{1}{k} \quad \forall j, l \in X_1 \quad (4.14)$$

$$a_m \leq b_j \quad \forall m \in X_0, j \in X_1 \quad (4.15)$$

$$a_m, b_j \geq 0 \quad \forall m \in X_0, j \in X_1 \quad (4.16)$$

The minimization problem is identical, except the signs with which the constraints enter the Lagrangean are reversed. Once again, all of the action is in the constraints and the complementary slackness conditions, rather than the objective function. Solution to both the minimization and maximization problem are guaranteed by the Weierstrass theorem. The intuitive arguments to reduce the number of cases under consideration are essentially the same as in the previous subsection. First order conditions and derivations are provided in the Appendix.

The solution to maximization problem (4.13)-(4.16) is as follows:

$$\begin{aligned} b_i^* &= \frac{2N_1-1}{kN_1} \\ b_j^* &= b_i^* - \frac{1}{k} = \frac{N_1-1}{kN_1} \quad \forall j \in X_1 \setminus i \\ a_l^* &= 0 \quad \forall l \in X_0 \end{aligned} \quad (4.17)$$

The solution to the similarly formulated minimization problem, for which b_i is the measure to be minimized, is given by:

$$\begin{aligned} b_i^* &= \frac{1}{kN} \\ b_j^* &= b_i^* + \frac{1}{k} = \frac{N+1}{kN} \quad \forall j \in X_1 \setminus i \\ a_l^* &= b_i^* = \frac{1}{kN} \quad \forall l \in X_0 \end{aligned} \tag{4.18}$$

Note that there is a more direct route to these answers, bypassing the Kuhn-Tucker problem, if we accept some further intuition. At the measure p that includes the upper bound on b_i , it is clear that all of the empty bins must contain measure 0. Furthermore, for this p , the final counter placed in any bin $j \in X_1 \setminus i$ *could* have been placed in b_i , but was not due to sheer randomization. This is tantamount to saying that if we were to remove the measure $\frac{1}{k}$ represented by the final counter in bin i , the residual allocation in bin i would still be exactly equal to that in all bins $j \in X_1 \setminus i$. That is, $b_i - \frac{1}{k} = b_j$ for all $j \in X_1 \setminus i$. Putting these insights together, we have $b_i + (N_1 - 1)(b_i - \frac{1}{k}) = P_R = \frac{N_1}{k}$, which rearranges to the solution above. An analogous line of argument provides the lower bound on b_i .

These results lead directly to Proposition 4:

Proposition 4. *Given any allocation $x \in \mathbb{Z}_+^N$ and the ensuing values $N_0, N_1 \in \mathbb{Z}_+$ and $P_R \in [0, 1]$, the probability measure in any bin $j \in X_1$ satisfies $p_j^* \in \frac{x_j - 1}{k} + [\frac{1}{Nk}, \frac{2N_1 - 1}{N_1 k}]$.*

Consider our standard example, with $x = (0, 1, 5, 4, 0)$. The seven counters in bins 3 and 4 that were not the final counter placed in their respective

bins account for $P_I = 0.7$, and we know that $p \geq \tilde{p} = (0, 0, 0.4, 0.3, 0)$. The maximum value of, for example, b_2 is that for which bin 2 could have received a second counter, but did not due to sheer randomization. This suggests that $b_2^{max} = b_j + \frac{1}{k}$ for $j = \{3, 4\}$. The corresponding p_2 -maximizing measure is $p = (0, 0.166, 0.466, 0.366, 0)$. Similarly, the minimum value of b_2 is $b_2^{min} = 0.02$, which gives the p_2 -minimizing measure $p = (0.02, 0.02, 0.52, 0.42, 0.02)$. The maximum and minimum values of b_3 and b_4 are the same as those for b_2 .

Width of the identified region

The bounds derived in the previous two subsections depend on N and k , characteristics of the beliefs elicitation module, and on N_1 , a characteristic of the response. Therefore the bounds vary across individuals. The full width of the identified intervals are easily calculated as the following:

1. Empty bin width: $\frac{N_1}{k(N_1+1)}$
2. Non-empty bin width: $\frac{1}{k}(2 - \frac{1}{N_1} - \frac{1}{N})$

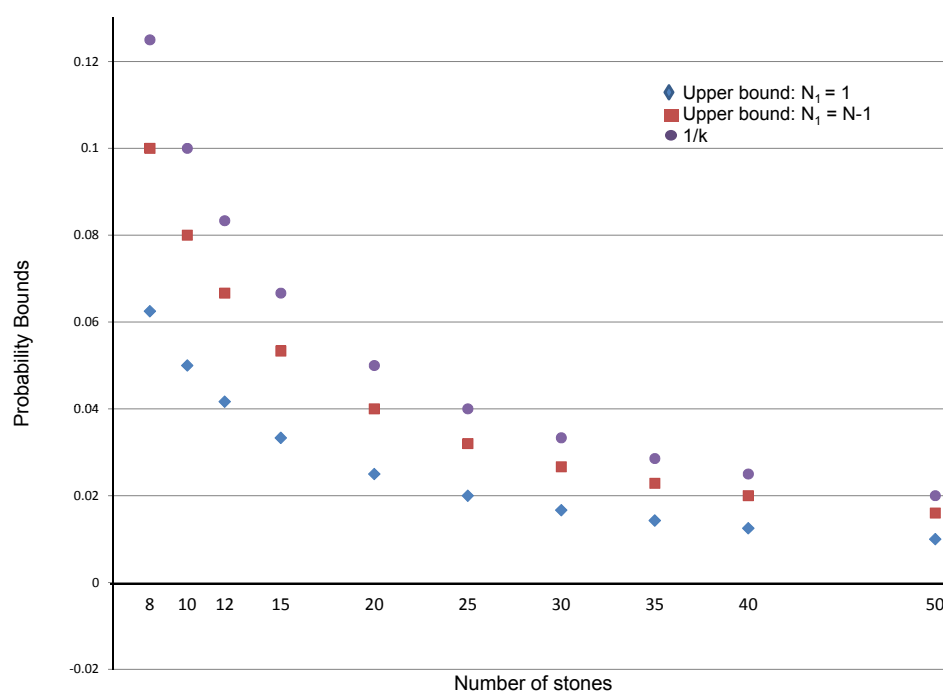
Note that the width of the identified interval for an empty bin is always less than $\frac{1}{k}$, the raw value of a counter. This is because the density in empty bin i is naturally bounded below by $\frac{x_i}{k} = 0$. Conversely, the width of the identified interval for a non-empty bin is always greater than or equal to $\frac{1}{k}$, unless there is only one non-empty bin (i.e., $N_1 = 1$).

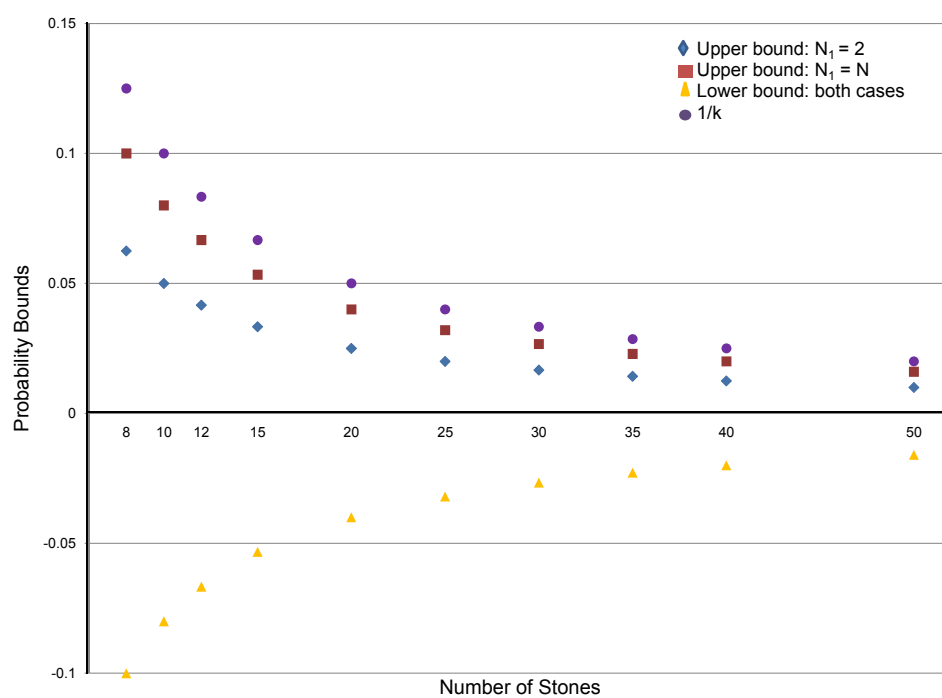
Figures 4.2 and 4.3 show the bounds of the identified regions, for empty and non-empty bins, respectively, holding $N = 5$ and allowing k to range from 8 to 50.¹³ Figure 4.2 depicts the upper bounds on the measure in an empty bin, for the cases $N_1 = 1$ and $N_1 = N - 1$. The value of $\frac{1}{k}$ is also depicted, for reference. Figure 4.3 shows the upper and lower bounds on the measure in any non-empty bin i , centered around $\frac{x_i}{k} \equiv 0$, for the cases $N_1 = 2$ and $N_1 = N$.¹⁴ The value of $\frac{1}{k}$ is also shown in Figure 3. Note that the lower bound of the residual measure in a non-empty bin is invariant to N_1 (see Proposition 4), so the series of lower bounds in Figure 4.3 applies to both $N_1 = 1$ and $N_1 = N - 1$.

It is clear in Figures 4.2 and 4.3 that as the number of bins receiving a counter increases, the identified region for any particular bin actually *increases* in size. This is yet another feature of these data that runs counter to our intuitions from working with sample data. We are accustomed to the notion that increased variation in any observed variable serves to better identify, for instance, the coefficients of a regression. However, with interdependent beliefs data, from the perspective of any bin i , allocation of the counters to a greater number of bins increases the total residual probability that might be occupied by the other bins, thereby increasing the range of values of p_i that are consistent with the observed allocation.

¹³In practice k is rarely above 20, but we include the higher values of k for sake of comparison.

¹⁴We begin with $N_1 = 2$ because the bounds on the non-empty bin measure are less interesting when $N_1 = 1$, because the upper bound is clearly $\frac{x_i}{k} = 1$.

Figure 4.2: Bounds on empty bin probabilities, $N = 5$

Figure 4.3: Bounds on non-empty bin probabilities, $N = 5$

4.3.2 Bounds on subsets of bins

Using the logical tools developed in the previous section, we now derive bounds on the total measure represented by any subset of the bins. We focus only on the maximization of the measure in any group of the bins, because any minimization problem can be written as the maximization of the measure in the complementary subset.

Let $\alpha \subseteq X_0$ be any subset of the set of empty bins, and $\beta \subseteq X_1$ be any subset of the set of non-empty bins. The problem of maximizing the total measure in the members of α and β can be written as:

$$\begin{aligned} \max_{(a_1, \dots, a_{N_0}, b_1, \dots, b_{N_1}) \in [0, \frac{1}{k}]^N} \sum_{i \in \alpha} a_i + \sum_{j \in \beta} b_j \quad \text{s.t.} \quad P_r &= \sum_{m=1}^{N_0} a_m + \sum_{j=1}^{N_1} b_j \quad (4.19) \\ b_j - b_l &\leq \frac{1}{k} \quad \forall j, l \in X_1 \\ a_m &\leq b_j \quad \forall m \in X_0, j \in X_1 \\ a_m, b_j &\geq 0 \quad \forall m \in X_0, j \in X_1 \end{aligned}$$

where the constraints are familiar from the previous subsection. Once again, the number of cases is extremely large, but intuition serves to reduce the problem to a few manageable possible solutions. The solution to any problem like (4.19) is unique, unless $\alpha = \emptyset$ and $\beta = X_1$.¹⁵ In any other case,

¹⁵There are infinite measures p , consistent with the observed x , for which the total measure in the empty bins is 0.

the solution to (4.19) is the unique vector satisfying:

$$(a_1, \dots, a_{N_0}, b_1, \dots, b_{N_1})^* \in [0, \frac{1}{k}]^N \text{ s.t. } \begin{cases} a_l = 0 & \forall l \in X_0 \setminus \alpha \\ a_i = b_m & \forall i \in \alpha, m \in X_1 \setminus \beta \\ b_j - b_m = \frac{1}{k} & \forall j \in \beta, m \in X_1 \setminus \beta \\ P_R = \sum_{m=1}^{N_0} a_m + \sum_{j=1}^{N_1} b_j \end{cases} \quad (4.20)$$

Letting A be the number of elements in α and B be the number of elements in β , some algebra reveals the solution to (4.19):

$$\begin{aligned} b_j^* &= \frac{2N_1 + A - B}{k(N_1 + A)} \quad \forall j \in \beta \\ a_i^* = b_m^* &= \frac{N_1 - B}{k(N_1 + A)} \quad \forall i \in \alpha, m \in X_1 \setminus \beta \\ a_l^* &= 0 \quad \forall l \in X_0 \setminus \alpha \end{aligned} \quad (4.21)$$

The intuition behind (4.20) is similar to that from the previous section. At the maximizing p , all a_i corresponding to $i \in \alpha$ must have measure exactly equal to that in $b_j \in X_1 \setminus \beta$, so that these bins could have received a counter, but did not due to sheer randomization. Otherwise, all $a_l \in X_0 \setminus \alpha$ are zero, and all $b_j \in \beta$ are $\frac{1}{k}$ greater than all $b_m \in X_1 \setminus \beta$.

Using the standard example of $x = (0, 1, 5, 4, 0)$, Table 4.2 gives the bounding residual probabilities for a variety of subsets α and β .

Table 4.2: Upper bounds on sum of bin measures, $x = (0, 1, 5, 4, 0)$, $P_R = 0.3$

α	β	Residual vector	Corresponding p	Bounding sum
$\{1, 5\}$	\emptyset	$(0.06, 0.06, 0.06, 0.06, 0.06)$	$(0.06, 0.06, 0.46, 0.36, 0.06)$	$p_1 + p_5 = 0.12$
\emptyset	$\{2, 3\}$	$(0, 0.133, 0.133, 0.033, 0)$	$(0, 0.133, 0.533, 0.333, 0)$	$p_2 + p_3 = 0.666$
$\{1\}$	$\{2\}$	$(0.05, 0.15, 0.05, 0.05, 0)$	$(0.05, 0.15, 0.45, 0.35, 0)$	$p_1 + p_2 = 0.2$
$\{1, 5\}$	$\{2\}$	$(0.04, 0.14, 0.04, 0.04, 0.04)$	$(0.04, 0.14, 0.44, 0.34, 0.04)$	$p_1 + p_2 + p_5 = 0.22$
$\{1\}$	$\{2, 3\}$	$(0.025, 0.125, 0.125, 0.025, 0)$	$(0.025, 0.125, 0.525, 0.325, 0)$	$p_1 + p_2 + p_3 = 0.675$
$\{1, 5\}$	$\{2, 3\}$	$(0.02, 0.12, 0.12, 0.02, 0.02)$	$(0.02, 0.12, 0.52, 0.32, 0.02)$	$p_1 + p_2 + p_3 + p_5 = 0.68$

Bounds on the CDF of $f(z)$

With (4.20) in hand, it is straightforward to bound the cumulative distribution function of underlying beliefs $f(z)$. Consider any $y \in \mathbb{R}$, and suppose that $\underline{d}_j < y \leq \bar{d}_j$, so that y lies in bin j . The maximum value of $\Pr(z < y)$ corresponds to a measure p that maximizes the sum of the density in bins $1 \dots j$. Bin j is included in the summation because the upper bound measure is one for which all of the density in bin j lies to the left of y , including, in the extreme case, the discrete distribution with location element $w_j = \underline{d}_j$. Similarly, the minimum value of $\Pr(z < y)$ corresponds to a measure p that maximizes the sum of the density in bins $j \dots N$. Bin j is included in this summation as well, because the bounding measure is one for which all of the density in bin j is assumed to lie to the right of y , including the discrete distribution with location element $w_j = \bar{d}_j$.

Consider the allocation $x = (0, 1, 5, 4, 0)$. To bound any y in bin 3, for example, we need only find the measure \underline{p} that maximizes $p_1 + p_2$, and the measure \bar{p} that maximizes $p_3 + p_4 + p_5$. Then it will be the case that

$\underline{p}_1 + \underline{p}_2 < (\Pr(z < y)) < \bar{p}_1 + \bar{p}_2 + \bar{p}_3$. Using (4.20), we can easily calculate the result: $0.025 < (\Pr(z < y)) < 0.675$.

For any bin i , let L_1^i and R_1^i be the number of non-empty bins to the left and right of bin i , respectively. Similarly, let L_0^i and R_0^i be the number of empty bins to the left and right of bin i , respectively. Propositions 5 and 6 give general characterizations of the bounds on the value of $\Pr(z < y)$, for any given $y \in \text{support}(z)$, for non-empty and empty bins, respectively.

Proposition 5. *Let $x \in \mathbb{Z}_+^N$ be an allocation with related values $N_0, N_1 \in \mathbb{Z}_+$, and let $y \in \mathbb{R}$ be any scalar lying in **non-empty** bin i , such that $\underline{d}_i < y < \bar{d}_i$. Then the value of $\Pr(z < y)$ corresponding to underlying beliefs $f(z)$ satisfies $\Pr(z < y) \in \frac{1}{k} \sum_{j=1}^{i-1} \max\{0, x_j - 1\} + \left[\frac{(L_1^i)^2}{k(N_1 + R_0^i)}, \frac{x_i}{k} + \frac{1}{k} \left(L_1^i + \frac{iR_1^i}{N_1 + L_0^i} \right) \right]$.*

Proposition 6. *Let $x \in \mathbb{Z}_+^N$ be an allocation with related values $N_0, N_1 \in \mathbb{Z}_+$, and let $y \in \mathbb{R}$ be any scalar lying in **empty** bin i , such that $\underline{d}_i < y < \bar{d}_i$. Then the value of $\Pr(z < y)$ corresponding to underlying beliefs $f(z)$ satisfies $\Pr(z < y) \in \frac{1}{k} \sum_{j=1}^{i-1} \max\{0, x_j - 1\} + \left[\frac{(L_1^i)^2}{k(N_1 + R_0^i + 1)}, \frac{1}{k} \left(L_1^i + \frac{iR_1^i}{N_1 + L_0^i + 1} \right) \right]$.*

Proofs of Propositions 5 and 6 are given in the Appendix. Note that for given values of N_1, L_1^i, R_1^i and L_0^i , the bounds on $\Pr(z < y)$ for y in an empty bin are always less than those for y in a non-empty bin.

Figures 4.4 and 4.5 show the bounds on the value of $\Pr(z < y)$, when underlying beliefs are $z \sim N(570, 110)$, for the cases of $N = 5$ and $N = 10$, respectively. The support of z is restricted to the interval $[0, 1000]$.¹⁶ We use this $f(z)$ because when $N = 5$, it leads to the allocation $x = (0, 1, 5, 4, 0)$, our

¹⁶To satisfy the assumption that the region of $f(z)$ with positive support lies entirely within the range of the bins, the normal distribution was truncated at 0 and 1000, and the total tail probability, $\int_{-\infty}^0 F(x)dx + \int_{1000}^{\infty} F(x)dx = 0.00004643$ was redistributed uniformly across the interval $[0, 1000]$.

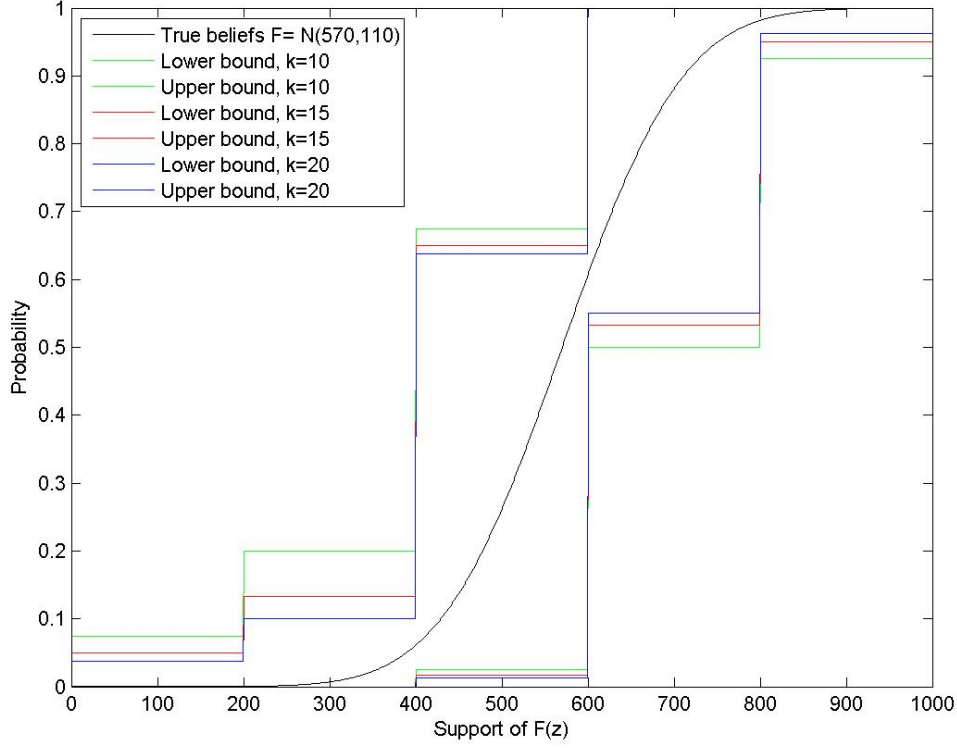


Figure 4.4: Bounds on $\Pr(z < y)$, $N = 5$, $f(z) = \mathcal{N}(570, 110)$

standard example. In both figures, bounds are shown for $k = \{10, 15, 20\}$. The bounds depicted do not themselves constitute CDFs that are consistent with the observed x , because the bounding p for y in bin i is inconsistent with the bounding p for y in any other bin $j \neq i$. Instead, the bounds shown are the outer envelopes of the 5 (in Figure 4.4) or 10 (in Figure 4.5) bounding CDFs that are consistent with the bounds in Propositions 5 and 6 for a y that falls in each of the N bins.

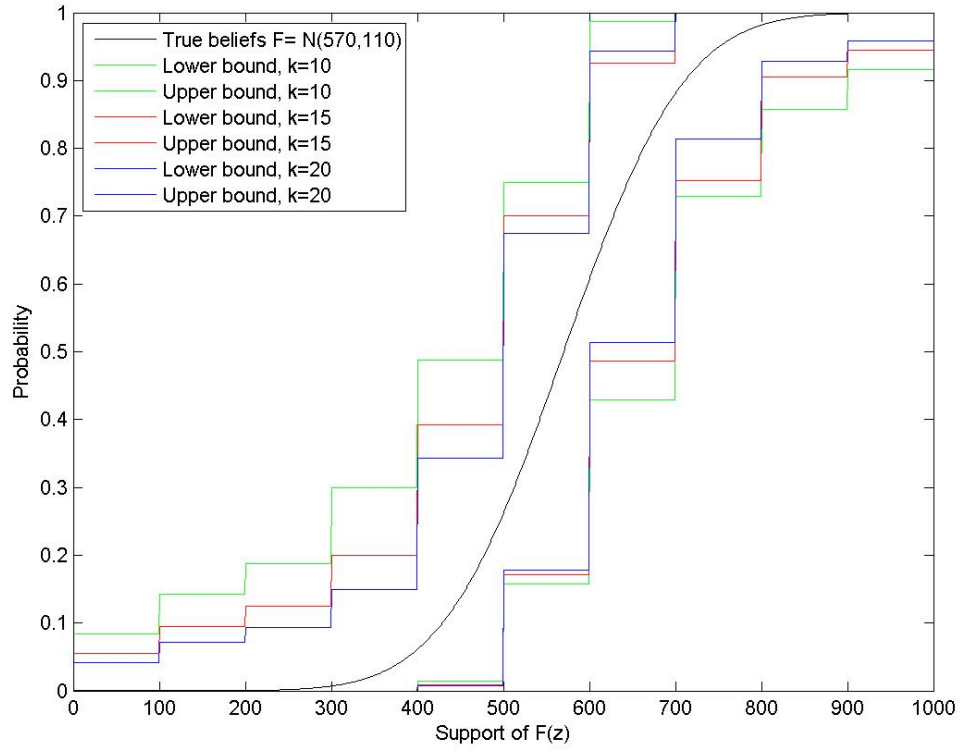


Figure 4.5: Bounds on $\Pr(z < y)$, $N = 10$, $f(z) = N(570, 110)$

4.3.3 Bounds on the median of f

Using (4.20) and (4.21), it is a simple task to bound the median M of $f(z)$. Recall that by definition, $F(M) = \int_{-\infty}^M f(z)dz = 0.5$. Then the lowest value of the median that is consistent with x is \underline{d}_j , where j is the leftmost bin such that $\max(\sum_{i=1}^j p_i) \geq 0.5$. This lower bound is quickly found by iteratively applying the results in (4.21) to subsets of bins $\{1\}, \{1, 2\}, \dots, \{1, 2, \dots, i\}$, until a bin is located that satisfies the criteria. By an exactly analogous argument, the upper bound on M is \bar{d}_l , where l is the rightmost bin such that $\min(\sum_{i=1}^l p_i) \leq 0.5$.

4.3.4 Joint Identification Region for p

The bounds (4.1)-(4.5) from Proposition 2 provide a means for characterizing the joint identification region of the elements of p , which we refer to as $\Phi \subset [0, 1]^N$. As is clear from the previous section, the extreme values of the measures in more than one bin are not mutually attainable in any p that is consistent with x .¹⁷ The identified set, therefore, will not be “rectangular” in the N -dimensional sense. Instead, it is a convex set characterized by the intersection of closed half spaces.

Result (4.5) implies that for all ij pairs, any $p \in \mathbb{R}^N$ that is in the identified set must lie in the lower half¹⁸ of the half-space ϕ_{ij} , defined as

¹⁷The exceptions are the minimum measures of 0 in the empty bins, which are jointly attainable.

¹⁸The side containing the origin $\mathbf{0} \in \mathbb{R}^N$.

$\phi_{ij} \equiv [q^{ij} \cdot p \leq \alpha^{ij}]$, for scalar α^{ij} and $1 \times N$ vector q^{ij} satisfying:

$$q^{ij} = \begin{bmatrix} 0 & \dots & 0 & 1 & 0 & \dots & 0 & -1 & 0 & \dots \end{bmatrix} \quad (4.22)$$

$i \qquad \qquad \qquad j$

$$\alpha^{ij} = \frac{x_i - x_j + 1}{k} \quad (4.23)$$

There are $N(N - 1)$ such ϕ_{ij} supporting vectors. However, only those for which bin j is non-empty are binding. When j is empty the lower bound $p_j \geq 0$ supersedes the restrictions imposed by (4.22)-(4.23). Therefore there are only $N_1(N - 1)$ half-spaces defined by the above that actually bound Φ . In addition, the lower bounds $p_l \geq 0$ on empty bins l can be represented by the N_0 half-spaces $[q^l \cdot p \leq 0]$ where $q_l^l = -1$ and $q_m^l = 0 \forall m \neq l$. The joint identification region Φ , then, is the intersection of these $N_0 + N_1(N - 1)$ sets.

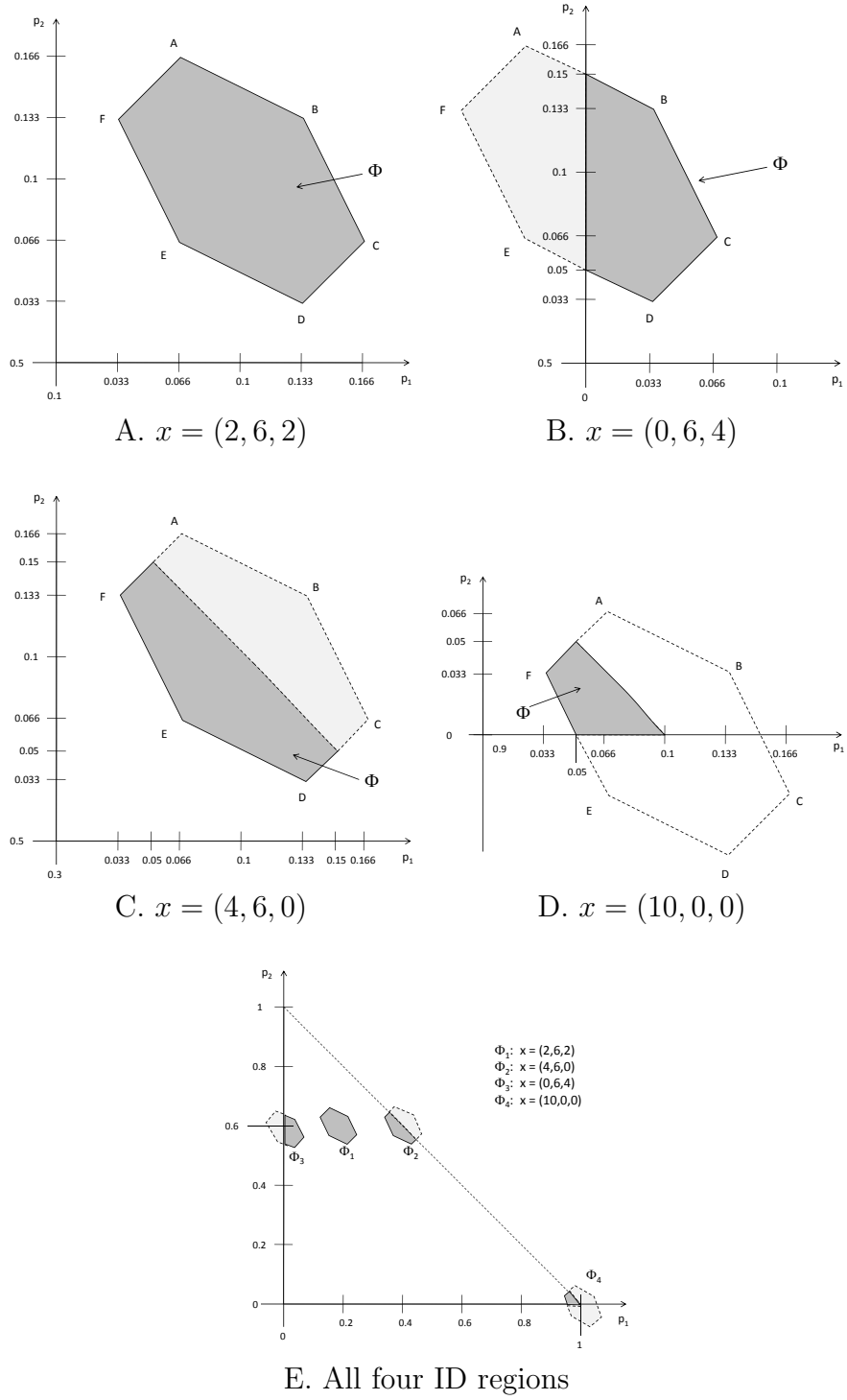
Consider some examples for the case $N = 3$, $k = 10$. The adding up constraint $\sum_{i=1}^N p_i$ allows us to represent an N -dimensional measure vector in $(N - 1)$ -space. Without loss of generality we exclude the 3rd bin.¹⁹ Figure 4.6 shows the identified region Φ for $x = (2, 6, 2)$, $x = (0, 6, 4)$, $x = (4, 6, 0)$ and $x = (10, 0, 0)$. The origin of each graph is $(\max\{0, \frac{x_1-1}{k}\}, \max\{0, \frac{x_2-1}{k}\})$. Because the restrictions embodied in (4.22) and (4.23) are linear in x and k , the polyhedron structure of the Φ in these figures is a general feature of subjective distributions data. For comparison, all four identified regions are

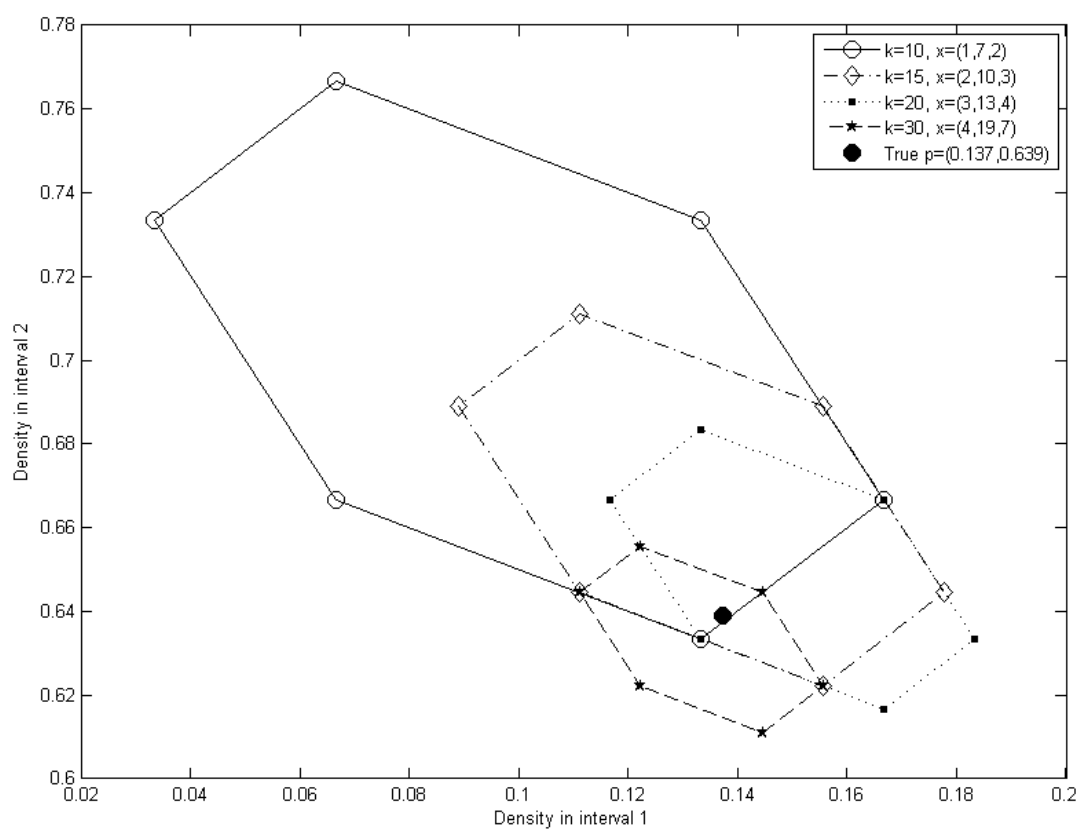
¹⁹The supporting vector for half-spaces that involve the excluded bin can be easily adjusted by substituting $1 - \sum_{i=1}^{N-1} p_N$ for p_N .

also displayed on a single set of axes, with the origin at $(0, 0)$.

There is no reasonable way to compare the size of the identified region as N increases, because the dimension of Φ changes with N . Comparisons across N would be analogous to comparisons between the area of a 2-dimensional figure and the volume of a 3-dimensional figure. However, we can study changes to the size of Φ as k increases, for a given N . Figure 4.7 gives an example for $N = 3$. For an underlying normal distribution with $(\mu, \sigma) = (5.3, 1.8)$, and bin bounds $d = (0, 3.33, 6.67, 10)$, the figure shows the joint identification region for 10, 15, 20 and 30 counters. The value of p_1 is on the horizontal axis, and p_2 is on the vertical axis. The true p is represented by a black dot. All 4 polyhedrons have the familiar structure. It is noteworthy that while the region corresponding to $k = 15$ is nested inside that corresponding to $k = 10$, the regions are not in general nested as k increases. This is a general feature of the problem as k increases with N fixed.

In Figure 4.7, the area of the identified region shrinks as k increases. This is generally the case, because the maximum difference between p_i and p_j , for any x and any bins i and j , is non-increasing in k . However, whenever the increase from k counters to $k + 1$ counters leads to an increase in N_1 , the size of the identified region Φ increases. This is because the researcher does not know which bin would not have received a counter if k were reduced by 1. Therefore, from an identification perspective, she cannot rule out the possibility that the marginal counter represents density greater than $\frac{1}{k}$. This is a counterintuitive finding, as we would not generally expect that providing


 Figure 4.6: Joint identification region Φ for various allocations

Figure 4.7: Identified region Φ as k increases

the respondent with more tools to convey information would actually reduce the amount of information conveyed.

Table 4.3 gives summary statistics for the percent decrease in the size of the identified region as k increases from 10 to 30, by 5, for $N = 3, 5$, and 10, for a sample of 200 underlying normal distributions.²⁰ Proportionally speaking, larger gains are observed for greater N , but this is partly due to the higher dimensionality of the problem. Unsurprisingly, the size of the gain decreases as k increases. Negative changes indicate that the size of the identified region actually increases with k . Such cases are more likely, and in general more severe, for small N . However, because we cannot make meaningful comparisons across N , the usefulness of Table 4.3 is limited. In the following section we will compare the size of the joint identified region for the moments of $f(z)$ as both N and k change, an exercise that will prove more useful for making recommendations about the choice of N and k .

The bounds that define Φ are sharp, in that they exhaust the distributional information provided by the respondent. Without additional information about the respondent's beliefs, all points in Φ are equally likely to be true representations of the underlying $f(z)$. Therefore, we can use Φ to bound the moments of $f(z)$, by considering the mean-variance pairs that are consistent with points in Φ and the boundaries of the N bins.

²⁰Exact volumes for the identified polytopes were not calculated, as the computational burden is substantial, particularly for the 10 bin case. Instead, we constructed grids in \mathbb{R}^N and counted the number of grid points inside Φ for each value of k .

Table 4.3: Percent change to size of Φ as k increases

k	N=3				N=5				N=10			
	Mean	SD	Min	Max	Mean	SD	Min	Max	Mean	SD	Min	Max
10	-	-	-	-	-	-	-	-	-	-	-	-
15	0.51	0.15	-0.31	0.57	0.76	0.11	0.35	0.82	0.90	0.08	0.50	0.98
20	0.41	0.15	-0.66	0.71	0.62	0.15	-0.08	0.82	0.85	0.08	0.39	0.94
25	0.33	0.13	-0.89	0.37	0.54	0.16	-0.31	0.85	0.76	0.15	-0.21	0.97
30	0.29	0.13	-1.10	0.33	0.50	0.16	-0.46	0.85	0.65	0.23	-0.22	0.91

Notes: mean is the average reduction in size of Φ when k is increased by 5; area approximated by counting number of grid points inside the joint identification region; grid step sizes = 0.02 for N=10, 0.0025 for N=5, 0.001 for N=3; author's calculation from 200 randomly selected underlying normal distributions, with the same 200 distributions used for each N

4.4 Bounds on the Expectation and Variance

In determining bounds on the moments of $f(z)$, the researcher faces a programming problem with choice vector (p, w) and constraints imposed by the boundaries of the bins and the requirement that $p \in \Phi$. While it turns out to be a straightforward task to bound the expectation by itself, the problem to bound the variance conditional on the mean taking a particular value, which is tantamount to jointly bounding the mean and variance, has no closed form analytical solution. In this section I first derive bounds on the mean for given allocation x and bin-boundary vectors \underline{d} and \bar{d} . I then characterize the problem of jointly bounding the mean and variance, prove a Lemma that simplifies the numerical algorithm to derive the joint identification region, and give examples of the joint identification region for particular cases.

4.4.1 Bounds on the expectation of f

It is intuitively clear that the minimum (maximum) expected value that is consistent with x will be the expectation a discrete distribution for which the density in all of the bins is stacked at the leftmost (rightmost) boundaries, given by the vector \underline{d} (\bar{d}). The minimizing measure vector is then the measure $p \in \Phi$ that minimizes $p'\underline{d}$. Given that the initial measure $z_i = \max\{0, \frac{x_i-1}{k}\}$ is already accounted for in each bin i , the contribution $z'\underline{d}$ to the mean is invariant to our choice of p . Therefore, the mean-minimization problem over p is equivalent to the following choice problem over the residual density vector $r = (a, b)$:

$$\min_{(a_1, \dots, a_{N_0}, b_1, \dots, b_{N_1}) \in [0, \frac{1}{k}]^N} \sum_{i \in X_0} a_i \underline{d}_i + \sum_{j \in X_1} b_j \underline{d}_j \quad \text{s.t. constraints (4.8)–(4.11) hold} \quad (4.24)$$

We can solve problem (4.24) in a manner that follows the spirit, if not the letter, of a proof by induction.²¹ Suppose that r^{i-1} is the residual vector that maximizes the measure in the leftmost $i-1$ bins. Let $m^0 = z'\underline{d} + r^{i-1'}\underline{d}$ denote the mean associated with this vector. Clearly any empty bins to the right of bin $i-1$ must have measure zero in r^{i-1} . We may then ask, under what circumstances is it optimal to re-allocate an ϵ -worth of density from each non-empty bin to the right of i , into bin i ? From the previous section and (4.20) we know that in vector r^{i-1} , $b_j - b_i = \frac{1}{k}$ and $a_m = b_i$ for

²¹One could also characterize the solution provided in this section as the solution of the dual linear programming problem to the primal given by equation (4.24).

all non-empty bins $j \leq i - 1$, empty bins $m \leq i - 1$ and non-empty bins $l > i - 1$. Therefore, if we move ϵ from each non-empty bin to the right of i into bin i , we also must move ϵ from each bin $1 \dots i - 1$ into bin i , in order to maintain consistency with the observed allocation x and satisfy the adding up constraint. Abusing notation slightly, let G_0^i and G_1^i indicate both the number of and the set of empty and non-empty bins, respectively, to the left ($G = L$) and right ($G = R$) of bin i (not including i itself). Then the mean m^1 after the proposed ϵ adjustment is given by:

$$m^1 = m^0 + \left[\underline{d}_i(L_1^i + L_0^i + R_1^i) - \left(\sum_{j \in L_1^i} \underline{d}_j + \sum_{l \in L_0^i} \underline{d}_l + \sum_{n \in R_1^i} \underline{d}_n \right) \right] \epsilon \quad (4.25)$$

which implies:

$$\begin{aligned} m^1 \leq m^0 &\iff \underline{d}_i(L_1^i + L_0^i + R_1^i) \leq \left(\sum_{j \in L_1^i} \underline{d}_j + \sum_{l \in L_0^i} \underline{d}_l + \sum_{n \in R_1^i} \underline{d}_n \right) \\ &\Rightarrow \underline{d}_i \leq \frac{\underline{d}_i + \sum_{j \in L_1^i} \underline{d}_j + \sum_{l \in L_0^i} \underline{d}_l + \sum_{n \in R_1^i} \underline{d}_n}{L_1^i + L_0^i + R_1^i + 1} = \underline{g}(i) \end{aligned} \quad (4.26)$$

Expression (4.26) suggests that for any level of ϵ , the proposed ϵ adjustment does not increase the mean of $f(z)$ if lower bound \underline{d}_i is less than or equal to the average of the lower bounds of all non-empty bins, all empty bins to the left of i , and \underline{d}_i itself. Therefore, one should re-allocate the maximum ϵ possible whenever this condition is satisfied. The maximum allowable

ϵ differs depending on whether i is empty or non-empty, as it is determined by the restrictions in (4.20) which maintain consistency with x . If i is empty, re-allocation should result in $a_i = b_l$, for a non-empty bin l to the right of i . Similarly, if i is non-empty, re-allocation should result in $a_i = b_j$ for non-empty bin j to the left of i . These constraints suggest the following choice of ϵ :

$$\begin{aligned} \text{If } i \text{ empty} & : & (L_1^i + L_0^i + R_1^i)\epsilon = (b_l - \epsilon) \Rightarrow \epsilon &= \frac{b_l}{L_1^i + L_0^i + R_1^i + 1} \\ \text{If } i \text{ non-empty} & : & (b_j - \frac{1}{k}) + (L_1^i + L_0^i + R_1^i)\epsilon = (b_j - \epsilon) \Rightarrow \epsilon &= \frac{1}{k(L_1^i + L_0^i + R_1^i + 1)} \end{aligned}$$

It is easy to confirm that in either case, this re-allocation results in r^i , the residual vector that maximizes the measure in the first i bins. Because our choice of i was general, these results suggest that the mean-minimizing vector will be associated with r^i , the vector that maximizes the density in bins $1 \dots i$, for the rightmost i that satisfies condition (4.26).

By an exactly analogous line of argument, the mean-maximizing vector will be associated with \hat{r}_l , the vector that maximizes the density in bins $l \dots N$, for the leftmost bin l that satisfies the following condition:

$$\bar{d}_l \geq \frac{\bar{d}_l + \sum_{j \in L_1^l} \bar{d}_j + \sum_{i \in R_0^l} \bar{d}_i + \sum_{n \in R_1^l} \bar{d}_n}{L_1^l + R_0^l + R_1^l + 1} = \bar{g}(l) \quad (4.27)$$

As an example, consider the allocation $x = (0, 1, 5, 4, 0)$ on the intervals bounded by $\underline{d} = (0, 2, 4, 6, 8)$ and $\bar{d} = (2, 4, 6, 8, 10)$. Table 4.4 gives the

Table 4.4: Bounding the expectation

Bin (i)	x_i	\underline{d}_i	\bar{d}_i	L_0^i	L_1^i	R_0^i	R_1^i	$\underline{g}(i)$	$\bar{g}(i)$	$\underline{d}_i \leq \underline{g}(i) ?$	$\bar{d}_i \geq \bar{g}(i) ?$
1	0	0	2	0	0	1	3	3	6	Yes	No
2	1	2	4	1	0	1	2	3	7	Yes	No
3	5	4	6	1	1	1	1	3	7	No	No
4	4	6	8	1	2	1	0	3	7	No	Yes
5	0	8	10	1	3	0	0	4	7	No	Yes

values relevant to the mean-minimization and mean-maximization problem for each bin:

The final two columns of Table 4.4 indicate that the minimum value of the mean that is consistent with x is associated with the residual vector that maximizes the density in the two leftmost bins, while the maximum value of the mean that is consistent with x is associated with the maximum density in the two rightmost bins. In the example given, the minimizing measure is $\underline{p} = (0.05, 0.15, 0.45, 0.35, 0)$, which gives a lower bound of $\underline{p}' \cdot \underline{d} = 4.2$. The maximizing measure is $\bar{p} = (0, 0.05, 0.45, 0.45, 0.05)$, giving an upper bound on the mean of $\bar{p}' \cdot \bar{d} = 7$.

4.4.2 Joint bounds on the expectation and variance

Rather than consider bounds on the variance in isolation, we now turn to the problem of jointly bounding the mean and the variance. It is straightforward to show that for any x and any feasible value of the mean, the maximum and minimum values of the variance that are consistent with that mean will be associated with discrete distributions. Formally, then, the problem to derive

an upper bound on the variance of $f(z)$, conditional on the mean taking value \bar{m} , can be written as:

$$\max_{p,w} \sum_{i=1}^N p_i \left(w_i - \sum_{i=1}^N p_i w_i \right)^2 \quad \text{s.t. } p \in \Phi \quad (4.28)$$

$$\begin{aligned} w_i &\in [\underline{d}_i, \bar{d}_i], \quad i = 1 \dots N \\ \sum_{i=1}^N p_i w_i &= \bar{m} \end{aligned} \quad (4.29)$$

The variance minimization problem is alike in all ways, except for the sign of the objective function. Because both p and w are choice variables, and the value of the mean inside parentheses varies with both, the objective function in (4.28) is non-convex, and closed form analytical solutions are not generally available. Unfortunately, even for relatively small values of N , numerical procedures to solve (4.28) can be very unstable and take prohibitively long to converge. Lemma 1, stated below, reduces the computational burden by placing additional restrictions on w at the solution to (4.28). This decreases the dimensionality of the search problem and substantially speeds up convergence.

Lemma 1. *Let $x \in \mathbb{Z}_N^+$ be an allocation, and let $p \in \Phi$ be any fixed measure vector that is consistent with x . Define an m -central location $w \in \mathbb{R}^N$ to be a location for which m elements of w , $1 \leq m \leq N$, satisfy $\underline{d}_i < w_i < \bar{d}_i$. Then the mean of any m -central location w , for $m > 1$, is also achieved by an $(m-1)$ -central location that has a greater variance, and by a different $(m-1)$ -central location that has a smaller variance.*

A proof of Lemma 1 is provided in the Appendix.²² The implication of Lemma 1 is that for a fixed measure p and a given value of the expectation of $f(z)$, the discrete distributions that achieve the minimum and maximum values of the variance are 1-central locations. Therefore, the only candidate solutions to (4.28) are pairs (p, w) that satisfy $p \in \Phi$ and $w_i \in \{\underline{d}_i, \bar{d}_i\}$ for at least $N - 1$ elements of the N -vector w .

We can further simplify the numerical procedure for solving (4.28) by using the mean constraint (4.29) and the adding up constraint $\sum_{i=1}^N p_i = 1$ to simplify the objective function. Note that if we satisfy (4.29) exactly, the term $(w_i - \sum_{i=1}^N p_i w_i)^2$ becomes a constant $v_i = (w_i - \bar{m})^2$ for each i . Substitution of v_i into (4.28) turns the problem into linear programming problem. In order to satisfy (4.29) exactly, we make the following substitutions for p_N and p_1 :

$$p_1 = 1 - \sum_{i=2}^N p_i \quad (4.30)$$

$$p_N = \frac{\bar{m} - \sum_{i=1}^{N-1} w_i p_i}{w_N} \quad (4.31)$$

For any given location vector w , (4.30) and (4.31) imply that $p_1 = \sum_{i=2}^{N-1} \gamma_i p_i$ and $p_N = \sum_{i=2}^{N-1} \eta_i p_i$ for fixed $N - 2$ dimensional vectors γ and η that are functions of w and \bar{m} . Substituting these expressions into the

²²Lemma 1 is also an implication of the main result proved separately in Stoye (2010) for a broad class of partial identification problems.

objective function (4.28), and also into the system of linear constraints that defines Φ , converts the conditional variance maximization (minimization) problem into a sequence of linear programming problems, one for each fixed w . Iterating over the set of location vectors w admitted by Lemma 1 leads to a solution. While this search procedure takes a substantial amount of time to converge for higher N (i.e. $N \geq 9$), the advantage is that the algorithms for solving linear programming problems are well understood, and we can be very confident that our solutions are global.

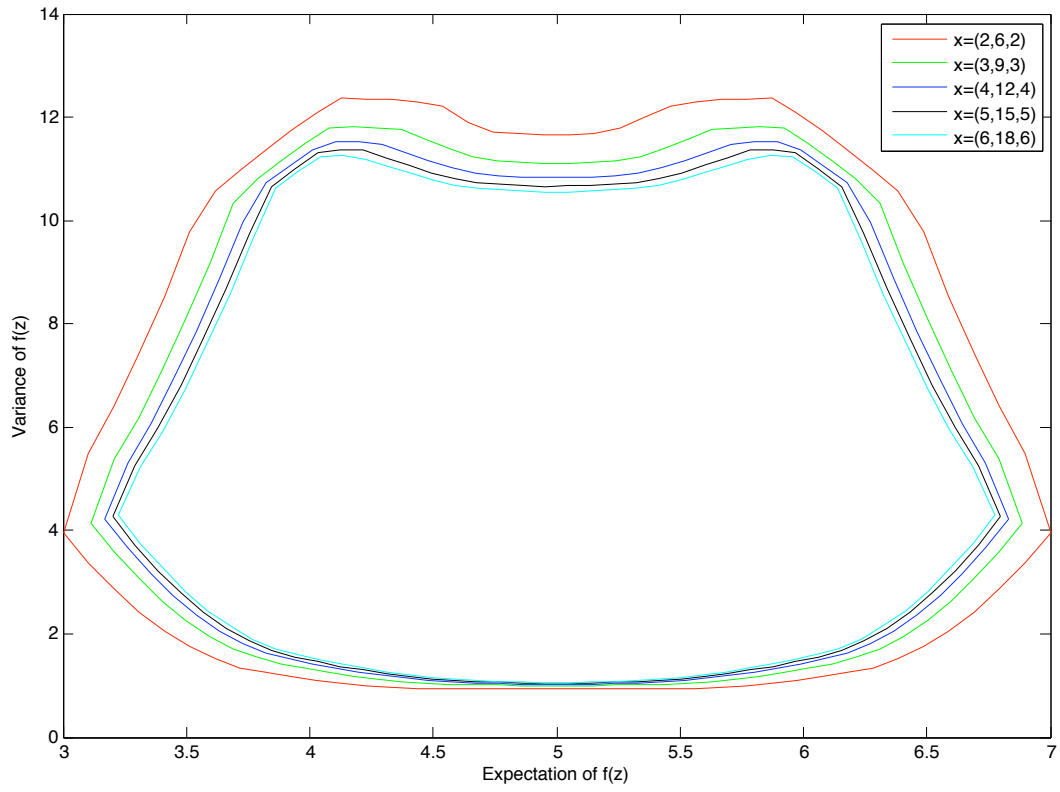
Figures 4.8 and 4.9 show the results for the two examples that we have considered most frequently in this paper. Figure 4.8 shows the joint identification region for $z \sim \mathcal{N}(5, 2)$, which induces the allocation $x = (2, 6, 2)$ when $k = 10$, for $k \in \{10, 15, 20, 25, 30\}$. The intervals are $d_1 = [0, \frac{10}{3}]$, $d_2 = [\frac{10}{3}, \frac{20}{3}]$, and $d_3 = [\frac{20}{3}, 10]$. There are a few things to note. First, as intuition would suggest, the identified region shrinks with the addition of more counters (we'll see momentarily that this is not always the case). The improvement is not dramatic, however. Second, substantial reductions in the range of the variance are only achieved as $E(z)$ approaches one of its boundaries. This suggests, unfortunately, that additional survey information, such as an elicited point expectation that can be used to select $E(z)$, may not substantially narrow the bounds on the variance.

Lastly, the identified regions are non-convex, with the non-convexity contributed by the upper bounds on the variance. Non-convex jointly identified mean-variance regions were frequently observed in the simulation exercise

described in the next section. The non-convexity appears to be driven by the fact that the variance is convex in the distance from the expectation. Toward the center of the support of the expectation, e.g., around $E(z) = 5$ in Figure 4.8, the maximum distance between the expectation and any point with positive density is at a minimum. As the expectation moves further from the midpoint of its identified interval, e.g., near $E(z) = 4$ or $E(z) = 6$ in Figure 4.8, the maximum distance between locations with positive density and the mean is greater than it is at the midpoint, and the bins containing these locations are still weighted with positive density. As the expectation approaches its extremes, e.g., near $E(z) = 3$ or $E(z) = 7$ in Figure 4.8, the maximum distance to points of positive support increases, but the total contribution of these farther points to the variance falls because the density in their associated intervals approaches its lower bound. In summary: the contribution of any particular bin to the variance is a function of the density in that bin and the distance between the location in that bin and the mean, and inter-play between these two forces determines the rise and then the fall in the variance.

Figure 4.9 shows similarly formulated identification regions, for $k = \{10, 15, 20, 25, 30\}$, bin boundaries that are the 5 evenly spaced bins between 0 and 10, and underlying true beliefs $z \sim \mathcal{N}(5.7, 1.1)$. These beliefs generate the allocation vector $x_1 = (0, 1, 5, 4, 0)$ when $k = 10$.

There is one key feature of Figure 4.9 that is deserving of attention. As k increases, the identified regions are not nested. This is counterintuitive, as

Figure 4.8: Mean-variance joint identification region, $f(z) = \mathcal{N}(5, 2)$, $N=3$

it would seem that the identification region should shrink in all directions as the respondent is able to represent her beliefs more finely. However, with the increase in counters from 15, the number of bins receiving a counter increases from 3 to 4. Even if the rightmost bin barely merits a counter, from an identification perspective this is unknown to the researcher. While this may seem like a stylized example, the general trend holds across other examples. Increasing the number of counters shrinks the mean-variance identification region as long as N_1 remains unchanged. The upper bound on the variance increases substantially when the marginal increase in k induces an increase in the number of non-empty bins.

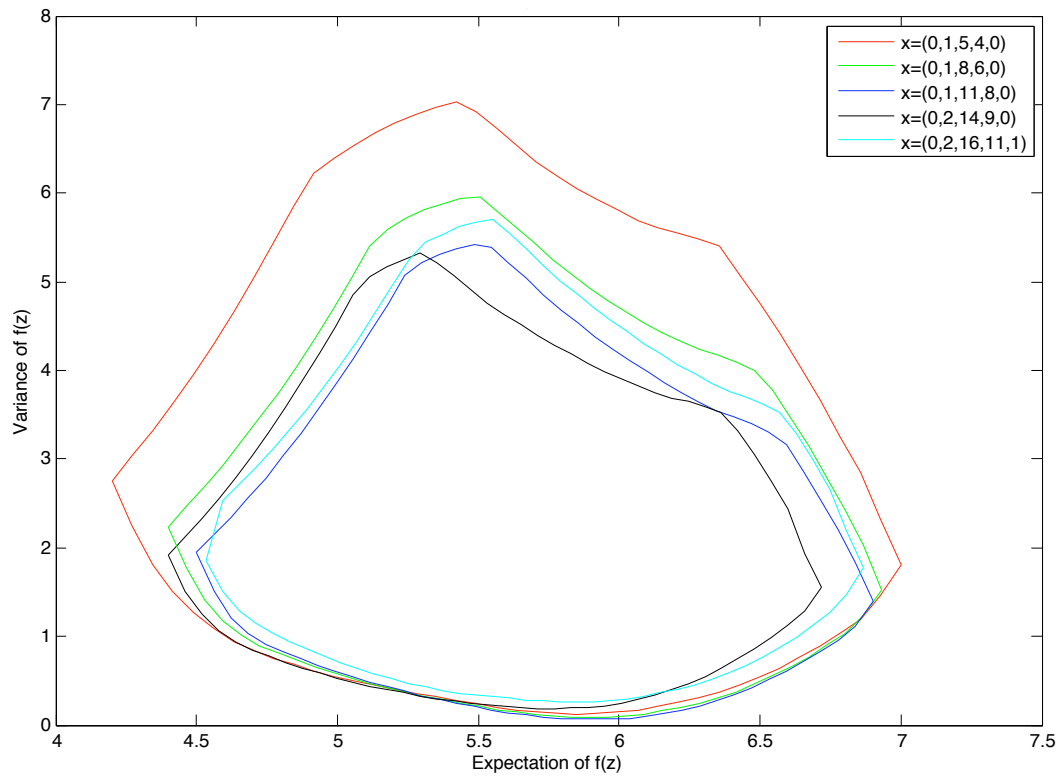


Figure 4.9: Mean-variance joint identification region, $f(z) = \mathcal{N}(5.7, 1.1)$, $N=5$

4.5 Simulation results

Delavande *et al* (2011) provide evidence from a field experiment that the moments of elicited distributions are not particularly sensitive to changes in the number of intervals or counters. However, the results in that paper rely on only a small number of combinations of N and k . Furthermore, the experiment participants were fishermen providing distributions over their daily catch, a random variable that is realized daily and with which they were all intimately familiar. Thus it is not surprising that these fishermen were able to communicate their subjective distributions consistently across survey modules.

In this section we provide results from two simulation exercises that inform the choice of N and k . First, we evaluate the effect of various choices of N and k on the size of the mean-variance identified region. Second, we consider the implications of various functional form choices on the estimation of moments from continuous approximations of x .

4.5.1 *Ex ante* choice of N and k

Results in this sub-section are based on a random sample of 500 (μ, σ) pairs drawn from the uniform intervals $\mu \in [2, 8]$, $\sigma \in [0.5, 2.5]$. For each (μ, σ) pair, we assume $f \sim N(\mu, \sigma)$ and calculate the measure vector p^N for $N \in \{3, \dots, 9\}$, where the N bins fully span the interval $[0, 10]$ with-

out overlapping.²³ We then generate the corresponding allocations for $k \in \{10, 15, 20, 25, 30\}$, and calculate the joint mean-variance identification region using the results of the previous section. Thus, for each of the 500 (μ, σ) pairs, we calculate $7 \times 5 = 35$ separate mean-variance regions corresponding to the unique combinations of N and k .

For comparison, we also calculate the mean-variance regions associated with the known measure vectors p^N . The vectors p^N are not observable to the researcher using the bin-and-counter method. However, one might argue that in surveys in the US such as the Survey of Economic Expectations, in which the respondent provides a numeric answer for the probability of z falling below fixed points $\{z_1, \dots, z_N\}$ on its support, the generated data is in fact p^N . For our purposes, the mean-variance regions based on the actual measure vector p^N provide the smallest achievable mean-variance region for a given N . Therefore, they give us a metric by which we can measure the relative change in the mean-variance region as k increases. If k were able to increase without bound, the mean-variance region from the bin-and-counter method would quickly approach that associated with p^N .

Tables 4.5, 4.6 and 4.7 show the simulation results. Table 4.5 shows the average area of the M-V region for combinations of N and k , as well as the limiting area for each N . There are two features of note. First, there is no simple answer to the question of whether it is better to increase N or k ,

²³We adjust the normal density function so that all of the density lies within the domain $[0, 10]$, by distributing the total density from outside this range uniformly over $[0, 10]$.

something that a researcher might struggle with during survey planning. If one were to begin in the box for $N = 3$, $k = 10$ and move iteratively toward the smaller area in either an easterly or southerly direction, one would begin heading south and then more or less alternate between increasing k and increasing N until reaching the lower right-most bin (ignoring the “limit” column). However, it is clear that small- N survey modules cannot effectively compensate by increasing the number of counters. The M-V areas associated with $N = 3$ are the largest of all the areas measured, regardless of k .²⁴ There are clear benefits to using at least 4 or 5 bins, before increasing k . Second, the limiting areas shrink more rapidly than do the identified regions, as N increases. Thus, in proportional terms, incremental increases in k provide less information for higher N than smaller N . Of course, to interpret this as a reason not to increase N would be backwards reasoning. Tables 4.6 and 4.7 show the changes in areas associated with rightward and downward movements across Table 4.5, respectively.

4.5.2 *Ex post* smoothing

In some situations, a researcher may wish to recover not just one or two moments from the elicited subjective distribution, but the entire distribution itself. A continuous approximation of $f(z)$ may be used in a structural model, or mixed with another distribution to approximate the distribution of the


²⁴We do not report results for $N = 2$, because the identified regions are so large that one can hardly consider such data to have any information content for the variance.

Table 4.5: Average area of M-V region for various (N, k) combinations

N	k					
	10	15	20	25	30	limit
3	29.54	26.06	24.32	23.25	22.55	18.91
4	19.87	16.79	15.24	14.33	13.64	10.33
5	16.23	13.06	11.41	10.52	9.90	6.68
6	14.45	11.13	9.45	8.46	7.86	4.68
7	13.60	10.03	8.27	7.25	6.58	3.43
8	13.22	9.47	7.61	6.52	5.83	2.63
9	13.06	9.07	7.18	6.07	5.35	2.08

Table 4.6: Percent change to size of $M - V$ region as k increases, N fixed

N	k				
	10	15	20	25	30
3	-	-11.5	-7.0	-4.6	-3.5
4	-	-15.4	-9.1	-6.1	-4.9
5	-	-19.4	-12.6	-7.8	-6.0
6	-	-22.7	-15.0	-10.3	-7.1
7	-	-25.9	-17.2	-12.2	-9.1
8	-	-28.1	-19.5	-14.0	-10.5
9	-	-30.5	-20.6	-15.2	-11.7



Entries are percent change in area of mean-variance identified region from box to the left

Table 4.7: Percent change to size of $M - V$ region as N increases, k fixed

N	k				
	10	15	20	25	30
3	-	-	-	-	-
4	-27.8	-29.8	-30.1	-29.8	-29.2
5	-16.6	-20.5	-23.4	-24.8	-25.6
6	-10.1	-14.2	-16.5	-18.7	-19.6
7	-5.7	-9.9	-12.4	-14.4	-16.3
8	-2.0	-5.3	-8.0	-9.9	-11.4
9	-0.3	-3.7	-5.3	-6.8	-8.0

Entries are percent change in area of mean-variance identified region from box above

product of two random variables. In this section I provide some initial Monte Carlo evidence for optimal smoothing methods. First, I discuss the procedure used to fit any general distribution function to an allocation x .

Suppose $g(z|\lambda)$ is a known cumulative distribution function, governed by parameter λ , that one wishes to fit to an allocation x . One natural method for doing so is to choose λ to minimize the square loss between $g(\bar{d}_i|\lambda)$ and the points (\bar{d}_i, q_i) , where $q_i = \sum_{j=1}^i p_j$. This choice problem can be written as:

$$\min_{\lambda} \sum_{i=1}^N [g(\bar{d}_i|\lambda) - q_i]^2 \quad (4.32)$$

This is the procedure advocated in Dominitz (1998) and used by a small number of recent papers. While the idea has intuitive appeal, it suffers from one primary shortcoming. If λ contains N_{λ} parameters, a solution to (4.32) is only identified for allocations with $N_1 > N_{\lambda}$. If $N = 5$ and one wishes

to fit a three-parameter distribution to the data, only responses for which 4 or 5 bins received a counter can be analyzed. A researcher can get around this problem by choosing extra points from within the identified region. But there is no optimal way to choose such points.

An alternative method for fitting g to x follows in the same vein as a bootstrap. The method is as follows: For very large Q , make Q draws from a uniform distribution on the interval $[\underline{d}_i, \bar{d}_i]$ for each counter in bin i . This generates kQ data points, which can then be fit to g using a simple GMM procedure or maximum likelihood. This bypasses any identification problems for the parameters of g , while remaining agnostic about the location of the density inside the bins. This procedure was used in the following analysis.

Table 4.8 shows the results from first ‘binning up’ known distributions into allocations, and then fitting various density functions to the binned data. Sample data was generated by drawing at random, from appropriately specified supports, 20 values of μ and 20 values of σ . The mean μ was drawn from the interval $[2.5, 7.5]$, and σ was drawn from $[0.5, 2.5]$. For each of the 400 resulting (μ, σ) pairs, the $\text{Normal}(\mu, \sigma)$ density function was then binned up to form an allocation x , following the minimization of absolute loss heuristic from Section 2. The bins were the 5 evenly spaced intervals between 0 and 10. This step was repeated for each of $k \in \{10, 15, 20, 30\}$, for the same fixed draw of 400 (μ, σ) pairs. Using the bootstrap method just described, with $Q = 2000$, these 400 allocations were then fit to the following known distributions: stepwise uniform, normal, generalized extreme value (“GEV”,

a 3-parameter distribution), generalized beta on $[0, 10]$, and generalized beta on $[\underline{d}_l, \bar{d}_m]$ where l is the first non-empty bin and m is the last non-empty bin. The normal and GEV distributions were fit using maximum likelihood, while the generalized beta distributions were fit using the method of moments.

The columns of Table 4.8 display the first significant digits of the following statistics: the square loss between the mean of the fit distribution and the mean of the true underlying distribution, the same for the standard deviation, the percentage of cases in which a particular distribution had the minimum total square loss for the first two moments, the area between the fitted and true CDFs, the percentage of cases in which a particular distribution had the minimum area between CDFs, and the Kolmogorov-Smirnov statistic. The bolded/highlighted entries in each “Best” column indicate the procedure that had the greatest likelihood of minimizing that particular measure of loss. Note the very poor performance of the stepwise uniform distribution. Though the sign of the bias isn’t shown, in nearly all cases the stepwise uniform distribution over-estimates the variance. This makes intuitive sense, and the variance is convex in the distance from the mean, and true beliefs are unlikely to be represented by a uniform step which falls suddenly to zero after passing some threshold. Thus, Table 4.8 provides some initial evidence suggesting that distributions other than stepwise uniform should be fit to observed allocations, when higher moments or full distributions are of interest to the researcher.

Table 4.8: Simulation results from *ex post* smoothing of known distributions

N	Fit Distribution	k=10						k=15					
		SQL mean	SQL sd	% SQL Best	Area btw CDFs	% Area Best	K-S stat	SQL mean	SQL sd	% SQL Best	Area btw CDFs	% Area Best	K-S stat
5	Stepwise Uni	0.034	0.071	0.282	0.285	0.068	0.080	0.023	0.065	0.260	0.255	0.087	0.078
	Normal	0.034	0.071	0.278	0.229	0.485	0.061	0.023	0.065	0.268	0.212	0.495	0.056
	Gen EV	0.034	0.074	0.44	0.243	0.305	0.070	0.022	0.066	0.472	0.225	0.282	0.068
	Gen Beta 1	0.034	0.071	-	0.254	0.078	0.080	0.023	0.065	-	0.241	0.052	0.077
	Gen Beta 2	0.034	0.071	-	0.273	0.072	0.089	0.023	0.065	-	0.254	0.108	0.086
8	Stepwise Uni	0.040	0.066	0.298	0.254	0.020	0.069	0.017	0.049	0.275	0.198	0.005	0.060
	Normal	0.040	0.066	0.260	0.211	0.438	0.055	0.017	0.049	0.213	0.162	0.440	0.042
	Gen EV	0.040	0.070	0.442	0.226	0.220	0.062	0.017	0.048	0.512	0.171	0.273	0.049
	Gen Beta 1	0.040	0.066	-	0.226	0.188	0.068	0.017	0.049	-	0.180	0.147	0.059
	Gen Beta 2	0.040	0.066	-	0.236	0.135	0.074	0.017	0.049	-	0.191	0.138	0.067
10	Stepwise Uni	0.045	0.079	0.380	0.263	0.020	0.071	0.028	0.038	0.248	0.209	0.013	0.059
	Normal	0.045	0.079	0.327	0.227	0.330	0.057	0.028	0.038	0.252	0.168	0.387	0.044
	Gen EV	0.044	0.085	0.292	0.232	0.257	0.059	0.027	0.037	0.500	0.177	0.317	0.050
	Gen Beta 1	0.045	0.079	-	0.241	0.162	0.068	0.028	0.038	-	0.187	0.207	0.059
	Gen Beta 2	0.045	0.079	-	0.240	0.230	0.071	0.028	0.038	-	0.199	0.080	0.065
N	Fit Distribution	k=20						k=30					
		SQL mean	SQL sd	% SQL Best	Area btw CDFs	% Area Best	K-S stat	SQL mean	SQL sd	% SQL Best	Area btw CDFs	% Area Best	K-S stat
5	Stepwise Uni	0.017	0.056	0.237	0.231	0.102	0.075	0.014	0.054	0.257	0.217	0.127	0.073
	Normal	0.017	0.056	0.235	0.188	0.550	0.051	0.014	0.054	0.273	0.179	0.558	0.049
	Gen EV	0.016	0.055	0.527	0.203	0.222	0.064	0.013	0.053	0.470	0.196	0.185	0.062
	Gen Beta 1	0.017	0.056	-	0.222	0.075	0.074	0.014	0.054	-	0.215	0.095	0.072
	Gen Beta 2	0.017	0.056	-	0.235	0.105	0.081	0.014	0.054	-	0.227	0.13	0.078
8	Stepwise Uni	0.015	0.033	0.282	0.173	0.030	0.055	0.009	0.028	0.310	0.144	0.060	0.053
	Normal	0.015	0.033	0.243	0.138	0.502	0.037	0.009	0.028	0.235	0.119	0.595	0.032
	Gen EV	0.014	0.031	0.475	0.152	0.237	0.046	0.008	0.025	0.455	0.136	0.175	0.043
	Gen Beta 1	0.015	0.033	-	0.161	0.162	0.056	0.009	0.028	-	0.148	0.093	0.054
	Gen Beta 2	0.015	0.033	-	0.173	0.085	0.062	0.009	0.028	-	0.157	0.117	0.059
10	Stepwise Uni	0.015	0.035	0.315	0.168	0.007	0.053	0.009	0.031	0.317	0.137	0.060	0.048
	Normal	0.015	0.035	0.257	0.137	0.450	0.034	0.009	0.031	0.255	0.117	0.512	0.029
	Gen EV	0.015	0.031	0.428	0.152	0.310	0.042	0.008	0.028	0.428	0.128	0.213	0.039
	Gen Beta 1	0.015	0.035	-	0.159	0.108	0.054	0.009	0.031	-	0.137	0.108	0.050
	Gen Beta 2	0.015	0.035	-	0.168	0.142	0.059	0.009	0.031	-	0.144	0.138	0.054

4.6 Conclusion

A rapid literature is emerging, primarily in development economics but also in labor and health economics, that attempts to better understand choice under uncertainty by measuring the subjective probability distributions of survey respondents. In this paper we have carefully considered the identification of the true underlying beliefs distribution when a particular method (the bin-and-counter method that has quickly become standard) is used to elicit distributions. We first analyzed the respondent's choice problem, and provided evidence on treating the data as the outcome of his attempt to minimize absolute loss between his belief and his response. This absolute loss heuristic has implications for the relationship between the true densities in any subset of the intervals in question. We exploit this fact to derive bounds on the measure in any single bin, bounds on subsets of bins, bounds on the CDF and the median, and the joint identification region for the measure in all of the bins. We also developed and tested a numerical procedure for estimating the jointly identified mean-variance region, and provided simulation evidence suggesting that small- N survey modules are particularly uninformative, and that stepwise uniform distributions are likely to fit the data worse than numerous other functional forms.

4.7 Appendix

Proof of Proposition 1

I provide three proofs, showing that placements of counters according to heuristics [A1.], [A2.] or [A3.] are consistent with minimization of absolute loss (subject to the restriction that all counters must be placed). This is sufficient to prove the equivalency of these heuristics. No proof is needed for heuristic [A4.], because it is defined with regard to the minimization of absolute loss, so the equivalency is trivial.

Proof. [A1.] We show that at each stage of iteration, placing the marginal counter in the bin suggested by heuristic [A1.] reduces absolute loss by at least as much as placement in any other bin. Define $l_i^s = |\frac{x_i^s}{k} - p_i|$ to be the contribution to the absolute loss from bin i at stage s , and note that $l_i^s = p_i^s$ if $p_i^s \neq 0$. Before placement of any counters, $x_i^0 = 0 \forall i \Rightarrow l_i^0 = p_i^0 \forall i \Rightarrow L^0 = 1$. Define $L^{s+1,i}$ to be the hypothetical loss in stage $s+1$ if the marginal counter in stage s is placed in bin i . Suppose that the respondent is at iterative stage $s \in \{0, \dots, k-1\}$. Let bin j be such that $p_j^s = \max_{i=1, \dots, N} \{p_i^s\}$. Heuristic [A1.] calls for the respondent to place the next counter in bin j . Let p_c^s be the measure in any other bin, with $p_c^s < p_j^s$. We consider three possible relationships between p_j^s, p_c^s , and $\frac{1}{k}$:

$$(i) \quad p_j^s, p_c^s > \frac{1}{k}.$$

Placement of the marginal counter in bin $j \Rightarrow L^{s+1,j} = \sum_{i \neq j} l_i^s + |p_j -$

$\frac{x_j^s+1}{k} = \sum_{i \neq j} l_i^s + (l_j^s - \frac{1}{k}) = L^s - \frac{1}{k}$, where we use the fact that $p_j^s > \frac{1}{k}$ implies $|p_j - \frac{x_j^s+1}{k}| = (p_j - \frac{x_j^s+1}{k})$. By the same sequence of arguments, placement of the counter in bin c gives $L^{s+1,c} = L^s - \frac{1}{k}$. So following [A1.] reduces absolute loss by no less than following any other heuristic.

$$(ii) \quad p_j^s > \frac{1}{k}, p_c^s < \frac{1}{k}.$$

Placement of the marginal counter in bin $j \Rightarrow L^{s+1,j} = L^s - \frac{1}{k}$ as above.

Placement of the marginal counter in bin c gives $L^{s+1,c} = \sum_{i \neq c} l_i^s + |p_c - \frac{x_c^s+1}{k}| = \sum_{i \neq c} l_i^s + |p_c^s - \frac{1}{k}| = \sum_{i \neq c} l_i^s + (\frac{1}{k} - p_c^s + (p_c^s - p_c^s)) = L^s + \frac{1}{k} - 2p_c^s$. Then $p_j^s > p_c^s \Rightarrow L^{s+1,j} < L^{s+1,c}$. Thus, following heuristic [A1.] reduces loss to a greater extent than does not following heuristic [A1.]

$$(ii) \quad p_j^s, p_c^s < \frac{1}{k}.$$

By the sequence of arguments in (ii), $L^{s+1,d} = L^s + \frac{1}{k} - 2p_d^s$ for $d \in \{c, j\}$. Therefore, $p_j^s > p_c^s \Rightarrow L^{s+1,j} < L^{s+1,c}$ once again.

We have shown that iteratively placing counters in the bins indicated by heuristic [A1.] always reduces absolute loss by at least as much as placing the counters in any other bins. This, together with the trivial observation that the absolute loss after all counters are allocated is invariant to the order of allocation, suffices to show that heuristic [A1.] minimizes absolute loss conditional on the requirement that all counters be allocated. \square

Proof. [A2.] We proceed in two steps: first, we show that placement of $kq_i = \text{floor}(p_i, \frac{1}{k})$, $i = 1, \dots, N$, in the initial stage is always feasible, and

that placement of fewer than kq_i counters in any bin is never consistent with minimization of absolute loss. We then show that during the “residual stage” the respondent allocates any remaining counters in a manner that minimizes absolute loss.

First, note that $\sum_{i=1}^N p_i = \sum_{i=1}^N (q_i + r_i) \Rightarrow k \sum_{i=1}^N q_i + k \sum_{i=1}^N r_i = k \Rightarrow \sum_{i=1}^N kq_i \leq k$. Therefore, during the initial stage, the respondent always has enough counters to place kq_i in bin i , $i = 1, \dots, N$. Denote by x^0 the initial stage allocation of counters. Suppose that in the initial stage the respondent places the requisite $k \sum_{i=1}^N q_i$ counters, but he places $x_j^0 < kq_j$ counters in bin j . Clearly a counter intended for bin j was placed in another bin, call it bin c . The contribution of bins j and c to the initial stage loss is $l_j^0 + l_c^0 = (p_j - \frac{x_j^0}{k}) + (\frac{x_c^0}{k} - p_c)$, where we use the under-allocation in bin j and over-allocation in bin c to identify the sign of these absolute value contributions. If only one such mistake was made then $x_c^0 = kq_c + 1$. In this case, re-allocation of one counter from bin c to bin j , using the superscript “1” to denote values after this re-allocation, gives the updated contribution to the loss: $l_j^1 + l_c^1 = (p_j - \frac{x_j^0}{k} - \frac{1}{k}) + (p_c - \frac{x_c^0}{k} + \frac{1}{k}) = (p_j - \frac{x_j^0}{k}) + (p_c - \frac{x_c^0}{k}) < l_j^0 + l_c^0$ because the second term is negative after cancellation of the $\frac{1}{k}$. If it happens that $x_c^0 > kq_c + 1$, which is the case if more than one extra counter was placed in c during the initial allocation, then by the same series of arguments it is easy to see that re-allocation of one counter from c to j reduces absolute loss by $\frac{2}{k}$. Therefore, failure to place less than kq_i counters in any bin i during the initial stage always increases absolute loss.

Second, note that as long as $r_i \neq 0$ for some i , some counters will be left over after the initial stage. Because $q_i \leq p_i \forall i$, the loss in all bins prior to the residual allocation is $p_i - \frac{x_i^0}{k} = p_i - r_i$. Heuristic [A2.] calls for allocation of any remaining counters to those bins i with the greatest values of r_i . Suppose that in the residual stage the respondent must choose whether to place a counter in bin k or in d , where $r_k > r_d$. If the residual counter is placed in bin k , total loss is $L^k = \sum_{i \neq k,c} l_i + (\frac{1}{k} - r_k) + (r_c)$, where L^k is the hypothetical loss if the counter is placed in bin k , borrowing notation from the previous proof. Likewise, $L^c = \sum_{i \neq k,c} l_i + (r_k) + (\frac{1}{k} - r_c)$. Clearly, $r_k > r_c \Rightarrow L^k < L^c$. So absolute loss is less if residual counters are placed in bins i with greater values of r_i .

We have shown that adherence to heuristic [A2.] never increases absolute loss, relative to non-adherence to [A2.], when all counters must be placed. Thus, the claim is proved. \square

Proof. [A3.] Note that by definition, $q_i \leq p_i = q_i + r_i < q_i + \frac{1}{k}$. Recall that $q_i = \text{floor}(p_i, \frac{1}{k}) \cdot \frac{1}{k}$, therefore q_i is divisible by $\frac{1}{k}$. For each bin i , rounding gives:

$$\bar{p}_i = \begin{cases} q_i & \text{if } r_i < \frac{1}{2k} \\ q_i + \frac{1}{k} & \text{if } r_i \geq \frac{1}{2k} \end{cases} \quad (4.33)$$

Let $D = \{i | \bar{p}_i = q_i\}$ and $U = \{i | \bar{p}_i = q_i + \frac{1}{k}\}$. These are the sets of rounded down and rounded up bins, respectively. As indicated in the statement of heuristic [A3.], the rounded measure need not sum to 1. We consider three

cases.

Suppose first that the rounded measure does sum to 1, i.e. $\sum_{i=1}^N \bar{p}_i = 1$. Heuristic [A3.] calls for allocation of $k\bar{p}_i$ counters to each bin i . Total absolute loss is $L = \sum_{i \in D} r_i + \sum_{j \in U} (\frac{1}{k} - r_j)$. Consider bin $c \in U$ and bin $d \in D$. By (4.33), it must be the case that $r_c > r_d$. Re-allocation of one counter from bin c to bin d gives $L' = \sum_{i \in D \setminus \{d\}} r_i + \sum_{j \in U \setminus \{c\}} (\frac{1}{k} - r_j) + r_c + (\frac{1}{k} - r_d) > L$. Therefore, such a re-allocation always raises absolute loss. Likewise, suppose a counter is re-allocated from bin $c \in U$ to another bin $b \in U$. The resulting total loss is $L'' = \sum_{i \in D} r_i + \sum_{j \in U \setminus \{b, c\}} (\frac{1}{k} - r_j) + r_b + (\frac{2}{k} - r_b) > L$. Similarly, if a counter is re-allocated from bin $d \in D$ to another bin $e \in D$, the resulting total loss is $L''' = \sum_{i \in D \setminus \{d, e\}} r_i + \sum_{j \in U} (\frac{1}{k} - r_j) + (\frac{1}{k} + r_d) + (\frac{1}{k} - r_e) > L$. So if $\sum_{i=1}^N \bar{p}_i = 1$, there is no re-allocation that does not raise absolute loss, relative to the allocation indicated by heuristic [A3.].

Next consider the situation in which $\sum_{i=1}^N \bar{p}_i < 1$. Heuristic [A3.] calls for initial allocation of $k\bar{p}_i$ counters to each bin. Denote total loss at this point by L^0 , and note $L^0 = \sum_{i \in D} r_i + \sum_{j \in U} (\frac{1}{k} - r_j)$. However, the allocation is not complete, because not all counters have been placed. Heuristic [A3.] calls for allocation of the k_r remaining counters to the k_r bins $i \in D$ with the greatest values of r_i (it is trivial to show that there must be at least k_r bins in set D). Suppose there are N_D bins in set D and N_U bins in set U . Let $r_D \in [0, \frac{1}{k})^{N_D}$ be the vector of residuals of the members of D , ordered from greatest to smallest. A typical element of r_D is r_D^a , where a is the rank of r_D^a among the residuals in D (1 being the greatest, N_D

the smallest). After placement of the k_r remaining counters, total loss is $L^1 = \sum_{j \in U} (\frac{1}{k} - r_j) + \sum_{b=1}^{k_r} (\frac{1}{k} - r_D^b) + \sum_{c=k_r+1}^{N_D} r_D^c$. Note that by definition of \bar{p}_i , $L^1 > L^0$. This is a case in which *unconditional* minimization of absolute loss would leave some counters unallocated. Suppose that instead of following the heuristic exactly, one of the k_r residual counters is placed in bin $e \in D$ s.t. $r_e = r_D^f < r_D^{k_r}$ instead of in bin $g \in D$ s.t. $r_g = r_D^h \geq r_D^{k_r}$. By definition, $r_e < r_g$. The resulting loss is $L^{1'} = L^1 - (r_e + (\frac{1}{k} - r_g)) + ((\frac{1}{k} - r_e) + r_g) = L^1 + 2(r_g - r_e) < L^1$. By a similar argument, it is easy to show that placement of one of the k_r remaining counters in bin $b \in U$, instead of bin $g \in D$, also increases absolute loss. So, conditional on the requirement that all counters be allocated, deviation from heuristic [A3.] cannot decrease absolute loss when $\sum_{i=1}^N \bar{p}_i < 1$.

Lastly, suppose $\sum_{i=1}^N \bar{p}_i < 1$. There are not enough counters to satisfy $x_i = k\bar{p}_i \forall i$. Treating the allocation of kq_i counters to each bin as the “initial stage”, and the allocation of the remaining counters as the “residual stage”, the series of arguments in the proof of heuristic [A2.] is directly applicable here.

Thus, for all possible values of $\sum_{i=1}^N \bar{p}_i$, we see that the Rounding heuristic is always consistent with minimization of absolute loss, subject to the requirement that all counters be allocated. \square

This completes the proof of Proposition 1.

Q.E.D.

Solution to Problem (4.8)-(4.11)

Although (4.8)-(4.11) is a straightforward linear programming problem with an analytical solution, it is subject to $N + N_0N_1 + N_1(N_1 - 1)$ inequality constraints, which means that there are $2^{(N+N_0N_1+N_1(N_1-1))}$ cases to consider. However, all of the non-empty residuals b_j can be treated symmetrically, as can all empty bin residuals other than residual a_i which appears in the objective function, $a_l \in A$ s.t. $l \neq i$. This still leaves 64 cases for consideration. We can use intuition to reduce the problem further. Clearly the maximization of a_i requires that all other empty bins have measure 0, so $a_i > 0$ and $a_l = 0 \forall l \neq i$. Likewise, for any $j \in X_1$, $b_j = 0$ only when $kp \in \mathbb{Z}_+^N$ and therefore $a_m = 0 \forall m \in X_0$, which will clearly not be the case when a_i is at a maximum. Therefore at a solution to (4.8), $b_j > 0$ and $b_j > a_l$ for all $j \in X_1, l \in X_0 \setminus i$. These insights dispense with the constraints in (4.11) and with all constraints in (4.10) except those that involve a_i . Only four unique cases remain, and of these four only one does not lead to a contradiction.

Proof of Proposition 5

Proof. Suppose that scalar y sits in non-empty bin i . Note that the density in bins $1 \dots i - 1$ that is accounted for by all but the final counter placed in each bin is $\frac{1}{k} \sum_{j=1}^{i-1} \max\{0, x_i - 1\}$. This is the minimum contribution to the value of $\Pr(z < y)$ for any y in bin i .

Recall from (4.20) that for the measure that maximizes the total density

in the members of subsets $\alpha \in X_0$ and $\beta \in X_1$, the following relationships must hold:

$$\begin{aligned} a_l &= 0 & \forall l \in X_0 \setminus \alpha \\ a_i &= b_m & \forall i \in \alpha, m \in X_1 \setminus \beta \\ b_j - b_m &= \frac{1}{k} & \forall j \in \beta, m \in X_1 \setminus \beta \\ P_R &= \sum_{m=1}^{N_0} a_m + \sum_{j=1}^{N_1} b_j \end{aligned}$$

The lower bound on $\Pr(z < y)$ corresponds to the measure which maximizes the total density in subsets $\alpha = R_0^i$ and $\beta = R_1^i \cup \{i\}$. Letting b_j be the density in any member of β , b_m be the density in any member of $X_1 \setminus \beta$, and a_l be the measure in any member of α , and using the fact that the residual measures must sum to P_R , we can write:

$$\begin{aligned} P_R &= (L_0^i)0 + (R_0^i)a_l + (L_1^i)b_m + (R_1^i + 1)b_j \\ \Rightarrow \frac{N_1}{k} &= (R_0^i + L_1^i)(b_j - \frac{1}{k}) + (R_1^i)b_j \quad \text{using (4.20)} \\ \Rightarrow b_j^* &= \frac{1}{k} + \frac{L_1^i}{k(N_1 + R_0^i)} \quad \text{and therefore} \quad a_i^* = b_m^* = \frac{L_1^i}{k(N_1 + R_0^i)} \end{aligned}$$

These collectively imply that r_i^{min} , the minimum contribution of bins $1 \dots i$ to the residual probability, is:

$$r_i^{min} = (L_1^i)b_m^* = \frac{(L_1^i)^2}{k(N_1 + R_0^i)}$$

Adding this to the initial density from bins $1 \dots i - 1$, we get the total

lower bound on the value of $\Pr(z < y)$ in non-empty bin i :

$$\underline{P}_R = \frac{1}{k} \sum_{j=1}^{i-1} \max\{0, x_j - 1\} + \frac{(L_1^i)^2}{k(N_1 + R_0^i)}$$

By a similar series of arguments, we can derive the upper bound on $\Pr(z < y)$ for any y in bin i . In this case let $\alpha = L_0^i$ and $\beta = L_1^i \cup \{i\}$. Then:

$$\begin{aligned} P_R &= (L_0^i)a_l + (R_0^i)0 + (L_1^i + 1)b_j + (R_1^i)b_m \\ \Rightarrow b_j^* &= \frac{1}{k} + \frac{R_1^i}{k(N_1 + L_0^i)} \quad \text{and therefore} \quad a_i^* = b_m^* = \frac{R_1^i}{k(N_1 + L_0^i)} \end{aligned}$$

which implies the following for r^{max} :

$$r_i^{max} = (L_0^i) \left(\frac{R_1^i}{k(N_1 + L_0^i)} \right) + (L_1^i + 1) \left(\frac{1}{k} + \frac{R_1^i}{k(N_1 + L_0^i)} \right) = \frac{1}{k} \left(L_1^i + 1 + \frac{iR_1^i}{N_1 + L_0^i} \right)$$

Adding this expression, and the value $\frac{x_i - 1}{k}$ which accounts for the initial density in bin i (which lies entirely to the left of y at the p that maximizes $\Pr(z < y)$), to the initial density in bins $1 \dots i - 1$ gives the upper bound on $\Pr(z < y)$:

$$\overline{P}_R = \frac{1}{k} \sum_{j=1}^{i-1} \max\{0, x_j - 1\} + \frac{x_i - 1}{k} + \frac{1}{k} \left(L_1^i + \frac{iR_1^i}{N_1 + L_0^i} \right)$$

□

Proof of Proposition 6

Proof. The structure of this proof is identical to that of the proof of Proposition 5. The lower bound on $\Pr(z < y)$ is derived as follows:

$$\begin{aligned}
P_R &= (L_0^i)0 + (R_0^i + 1)a_l + (L_1^i)b_m + (R_1^i)b_j \\
\Rightarrow b_j^* &= \frac{1}{k} + \frac{L_1^i}{k(N_1 + R_0^i + 1)} \quad \text{and therefore} \quad a_i^* = b_m^* = \frac{L_1^i}{k(N_1 + R_0^i + 1)} \\
\Rightarrow r^{min} &= (L_1^i)b_m^* = \frac{(L_1^i)^2}{k(N_1 + R_0^i + 1)} \\
\Rightarrow \underline{P_R} &= \frac{1}{k} \sum_{j=1}^{i-1} \max\{0, x_j - 1\} + \frac{(L_1^i)^2}{k(N_1 + R_0^i + 1)}
\end{aligned}$$

which is identical to the lower bound for a non-empty bin. This is because in both cases we assume that any density in bin i lies to the right of y , and therefore does not contribute to the minimum value of $\Pr(z < y)$.

Similarly, the upper bound on $\Pr(z < y)$ for y in empty bin i is derived as follows:

$$\begin{aligned}
P_R &= (L_0^i + 1)a_l + (R_0^i)0 + (L_1^i)b_j + (R_1^i)b_m \\
\Rightarrow b_j^* &= \frac{1}{k} + \frac{R_1^i}{k(N_1 + L_0^i + 1)} \quad \text{and therefore} \quad a_i^* = b_m^* = \frac{R_1^i}{k(N_1 + L_0^i + 1)} \\
\Rightarrow r^{max} &= (L_1^i)b_j^* + (L_0^i + 1)a_l^* = \frac{L_1^i}{k} + \frac{iR_i}{k(N_1 + L_0^i + 1)} \\
\Rightarrow \overline{P_R} &= \frac{1}{k} \sum_{j=1}^{i-1} \max\{0, x_j - 1\} + \left(\frac{L_1^i}{k} + \frac{iR_i}{k(N_1 + L_0^i + 1)} \right)
\end{aligned}$$

□

Proof of Lemma 1

Suppose $w \in \mathbb{R}^N$ is an m -central location, $m > 1$, associated with a fixed p that is consistent with an observed x . Without loss of generality, name two of the central elements 1 and 2, i.e. assume $\underline{d}_1 < w_1 < \bar{d}_1$ and $\underline{d}_2 < w_2 < \bar{d}_2$. We will show that the mean of the underlying distribution F , given by $\hat{m} = pw$, is achievable by at least two vectors for which $w_3 \dots w_N$ are unchanged, and either $w_1 = w'_1 \in \{\underline{d}_1, \bar{d}_1\}$ or $w_2 = w'_2 \in \{\underline{d}_2, \bar{d}_2\}$ (or both). The contribution of w_1 and w_2 to the mean is $p_1 w_1 + p_2 w_2$. Consider four mean-preserving changes to w (only two of which will generally be feasible, except at the boundary):

- (i) $w'_1 = \underline{d}_1$, $w'_2 = w_2 + \delta \in (w_2, \bar{d}_2]$, s.t. $p_1(w_1 - \underline{d}_1) = p_2(w'_2 - w_2)$
- (ii) $w'_1 = \bar{d}_1$, $w'_2 = w_2 - \epsilon \in [\underline{d}_2, w_2)$, s.t. $p_1(\bar{d}_1 - w_1) = p_2(w_2 - w'_2)$
- (iii) $w'_1 = w_1 + \delta \in (w_1, \bar{d}_1]$, $w'_2 = \underline{d}_2$, s.t. $p_1(w'_1 - w_1) = p_2(w_2 - \underline{d}_2)$
- (iv) $w'_1 = w_1 - \epsilon \in [\underline{d}_1, w_1]$, $w'_2 = \bar{d}_2$, s.t. $p_1(w_1 - w'_1) = p_2(\bar{d}_2 - w_2)$

Note that (i) is feasible iff. $\frac{p_1}{p_2} \leq \frac{\bar{d}_2 - w_2}{w_1 - \underline{d}_1}$ and (iv) is feasible iff. $\frac{p_1}{p_2} \geq \frac{\bar{d}_2 - w_2}{w_1 - \underline{d}_1}$, so that $\sim (i) \iff (iv)$ unless the equality holds, in which case both are true. Symmetrically, either (ii) or (iii) is true but not both, except at the boundary where both hold. This proves the feasibility in all cases of at least two different mean-preserving changes that replace m -central w with $(m-1)$ -central w' . That is, at least one of the following four pairs of mean-preserving changes is feasible in all cases: (i) and (ii), (i) and (iii), (ii) and (iv), or (iii)

and (iv).

It remains to show that for any pair of changes, whenever one change decreases the variance, the other necessarily increases it. Because $p, w_3 \dots w_N$ and \hat{m} are not changed by any of (i) – (iv), the contribution to the variance of all elements other than 1 and 2 is constant. The contribution of elements 1 and 2 to the variance, prior to any change, is $v_0 = p_1(w_1 - \hat{m})^2 + p_2(w_2 - \hat{m})^2$. Suppose first that (i) and (ii) are true. After the mean-preserving change in (i), the contribution of elements 1 and 2 to the variance of w' is

$$\begin{aligned}
 v_1^i &= p_1(\underline{d}_1 - \hat{m})^2 + p_2(w_2 + \delta - \hat{m})^2 \\
 &= p_1(w_1 - \hat{m} - (w_1 - \underline{d}_1))^2 + p_2(w_2 - \hat{m} + \delta)^2 \\
 &= v_0 + p_1(w_1 - \underline{d}_1) [(w_1 - \underline{d}_1 + \delta) + 2(x_2 - x_1)] \\
 \Rightarrow v_1^i > v_0 &\iff \frac{p_1 + p_2}{p_2}(w_1 - \underline{d}_1) > 2(w_1 - w_2) \quad (4.34)
 \end{aligned}$$

where we make use of the fact that $p_1(w_1 - \underline{d}_1) = p_2\delta$. By the same method it can be shown for v_1^{ii} , the contribution to the variance after (ii), that $v_1^{ii} > v_0 \iff \frac{p_1 + p_2}{p_2}(\bar{d}_1 - w_1) > 2(w_2 - w_1)$. It cannot be the case that $w_1 = w_2$. If $w_1 > w_2$, then $v_1^{ii} > v_0$ follows immediately. In addition, under the assumptions that (i) is true and $w_1 > w_2$, the condition for $v_1^i > v_0$ cannot hold, because:

$$\begin{aligned}
 (i) \Rightarrow \frac{p_1}{p_2} &\leq \frac{\bar{d}_2 - w_2}{w_1 - \underline{d}_1} \Rightarrow \frac{p_1 + p_2}{p_2}(w_1 - \underline{d}_1) \leq (\bar{d}_2 - w_2) + (w_1 - \underline{d}_1) \\
 &\Rightarrow \frac{p_1 + p_2}{p_2}(w_1 - \underline{d}_1) < 2(w_1 - w_2) \quad (4.35)
 \end{aligned}$$

which contradicts (4.34). We have shown that if (i), (ii) and $w_1 > w_2$ are true, then $v_1^{ii} > v_0$ and $v_1^i < v_0$. Symmetrical arguments show that (i), (ii) and $w_2 > w_1$ collectively imply $v_1^{ii} < v_0$ and $v_1^i > v_0$. And the same set of arguments hold if we begin with any of the other pairs of possible mean-preserving changes: (i) and (iii), (ii) and (iv), or (iii) and (iv). So in all cases, a pair of mean-preserving changes to w are feasible, both of which result in an $(m-1)$ -central location, and one of which increases the variance while the other decreases it. QED.

Chapter 5

Estimation of a Dynamic Agricultural Production Model with Observed, Subjective Distributions

5.1 Introduction

Although the technology of agricultural production in Sub-Saharan Africa is rudimentary by Western standards, depending as it does on rainfall, simple hand tools, and animal and human labor, the task of producing crops in an atmosphere of substantial risk and uncertainty is in fact a highly complex stochastic control problem. Farmers' solutions to this problem, which are

conditioned on subjective beliefs about prices, rainfall, and other uncertain agro-climatic conditions, are expressed as a sequence of crop acreage, labor allocation and non-labor input decisions. These decisions are usually made without reliable price and weather information. Methods to mitigate risk (such as crop insurance) and uncertainty (such as forward contracting) are rare. In such a setting, the design of appropriate policies to support wellbeing depends on the capacity of researchers to understand the dynamic choice problem of farmers.

This paper explores the roles of risk and uncertainty in the production decisions of cotton farmers in Tanzania. Our primary aim is to develop a method for incorporating subjective yield distributions into the estimation of a dynamic resource allocation model, so as to avoid making the untestable and largely unsubstantiated assumptions of rational expectations over a common distribution of production shocks. We use a unique, high-frequency panel data set of 195 farmers, who were surveyed once every three weeks over the entire course of the 2009-2010 cotton cultivation season. The data set includes regular measures of subjective price and yield distributions. That is, at regular intervals from planting to harvest, sample farmers used a visual aid to express their subjective belief about the full distributions of harvest-period yields and prices. Input allocation, time use, and other relevant data were also gathered regularly. Using this unique data set, we estimate a stochastic production model that incorporates the gradual revelation of information and the sequential nature of farmers' input choices.

This paper draws on, and contributes to, two lines of literature. First is the literature on farmer choice and dynamic decision-making under uncertainty. It is well understood that agricultural production is a sequential process in which farmers choose not only the level but also the timing of inputs, in accordance with their expectations and with the gradual resolution of uncertainty due to agro-climatic factors (Nerlove and Bessler, 2001). However, the general lack of data on the timing of inputs, and the near complete absence of data on farmer-level expectations over stochastic contributions to crop growth and revenue, has generally precluded estimation of decision models that take seriously the dynamic nature of the input allocation problem and the gradual resolution of uncertainty (Just and Pope, 2001).

Despite the prominent role of agriculture in sustaining the livelihoods of the rural poor in most developing countries, and despite the critical role that uncertainty plays in agricultural production, this literature has been stalled for almost two decades. The forebears of the model in this paper are the models in Hartley (1983), Wolpin (1987), and Fafchamps (1993). The estimation procedure in Fafchamps (1993), which identifies the inter-temporal elasticity of labor substitution from the sequential labor decisions of African farmers, is most similar to the method used in this paper. However, there are two key differences between the papers: first, we have measures of subjective yield and price distributions, so we do not need to make any assumptions about the parameters of the error distribution, nor do we need to restrict the stochastic component of output to be invariant across plots or across time.

Second, we explicitly allow for farmers to learn about the state of the world gradually, so that input choices depend on an assessment of crop progress to date, expectations about future stochastic contributions to output, and expectations about farmers' own future input choices. While the model in Fafchamps (1993) is sequential in a mechanical fashion, the estimation procedure treats the data as if all labor choices were made simultaneously.

The second line of literature related to this paper is the recent, rapidly growing body of work that utilizes measured, subjective expectations to study choice problems under uncertainty. Manski (2004) lays out the case for gathering data of this type, and Delavande *et al* (2010) provide a comprehensive review of recent work in this area. In a variety of developing country settings, researchers are using subjective expectations and distributions to study a range of choice problems, especially related to health and education. Notably lacking, however, are papers that use observed subjective distributions directly in the estimation of economic decision models. The one exception that we know of is Nyarko and Schotter (2002), in which subjective expectations measured in a lab experiment are used to estimate the parameters of a reinforcement model (and shown to out-perform models of expectations formation that are imposed on the choice data). To our knowledge, this paper represents the first attempt to use a sequence of observed subjective beliefs and contemporaneous choice data to estimate the parameters of a dynamic resource allocation model.

The paper proceeds as follows. In the next section we describe the data

set and the setting. In Section 5.3 we develop the stochastic production model, paying particular attention to the timing of information revelation, input allocation and data collection. In Section 5.4 we describe how the sequence of subjective output distributions reveals plot-level distributions of the stochastic contributions to output, and outline the estimation algorithm. Section 5.5 contains the results, and Section 5.6 concludes.

5.2 Data and Setting

In Tanzania, cotton is planted in late November and December, and harvested in May or June. The marketing season begins in late June and runs through September, with the large majority of sales taking place in July and August. In any given season, 30-50 private ginning companies compete to buy cotton through village-level buying agents. These companies gin the cotton, and export the lint. The government is not directly involved in the production, purchase or export of cotton, although the Tanzania Cotton Board (TCB) sets the opening date of the marketing season and broadcasts a minimum price that is calculated from the current world market price, building in a margin to cover the ginning companies' costs. The TCB minimum price is the modal transaction price during the first weeks of the marketing season. As the season progresses, ginning companies typically bid the price up as they push to meet their target purchasing volumes. Tanzania contributes only a fraction of total world cotton production, and Tanzanian cotton farmers are

exclusively small-scale. In addition, the price paid to farmers does not vary with the quality of the crop. Farmers, therefore, act as pure price-takers in the output market.

The data for this project were gathered over a one-year period from a sample of 195 cotton farmers in 15 villages in northwest Tanzania. Extensive face-to-face surveys were conducted with each farmer during a July-September 2009 baseline visit and a July-August 2010 follow-up. Baseline and follow-up surveys covered a wide range of standard LSMS survey topics. From September 2009 to June 2010, each sample farmer was interviewed once every 3 weeks on a prearranged schedule, for a total of 14 high frequency survey rounds.¹ These interviews covered cultivation activities for cotton and non-cotton plots, time use and borrowing, shocks, investment, other relevant economic and demographic data, and subjective probability distributions over end-season prices, end-season yields, pest pressure in the coming months, and rainfall levels in the coming months.

We gathered subjective probability distributions in a manner that has quickly become standard in development economics, by asking respondents to allocate a fixed number of counters to boxes that represent the bins of a histogram. The support of the price distribution, in units of Tanzania shillings per kilogram of seed cotton, was divided into seven intervals: Under

¹High frequency interviews were conducted using mobile phones that the team distributed during the baseline. For details on the phone-based survey method, see Chapter 3.

340, 340-400, 400-460, 460-520, 520-580, 580-640, Over 640.² Respondents were given 14 counters to allocate among the 7 price intervals. The yield support, in units of *mafurushi* per acre,³ was divided into 10 intervals labeled 0-9. Respondents were given 20 counters to allocate among the 10 yield intervals.⁴ Throughout the paper we assume that the price and yield density function are stepwise uniform, with the density in a given interval determined by the proportion of counters allocated to that interval.⁵

Table 5.1 gives household level summary statistics for sample households. Households are large: the mean size is 8.33 members, with the largest household containing 23 members. There is an average of 1.31 dependents (seniors and children) for every working age adult. The average age of heads of household is 46.85, and 85% of household heads are male. The mean education level of household heads is 4.19 years of schooling. Livestock holdings consist primarily of cows and goats, with the average household owning approximately 5 of each. There are 1.2 bicycles and 0.83 radios per household in the sample. During the 2008-2009 growing season, which began with the short rainy season in November of 2008 and ended in the summer of 2009, the

²The modal price from the previous (2009) marketing season was 440 TSH/kg.

³1 *mafurushi* \approx 90 kg.

⁴Subjective output distributions are calculated from subjective yield distributions by simply multiplying the interval boundaries by the number of acres. I refer to these two distributions interchangeably throughout the paper.

⁵In Chapter 4 we showed that the stepwise uniform assumption is almost never the optimal distributional assumption; nevertheless we make use of it in this chapter because it allows us to rapidly evaluate density functions as we develop the estimation technique with observed subjective shock distributions. In future work we relax the stepwise uniform assumption.

average farmer grew 3.45 crops on 9.7 acres spread across 2.7 plots. Other than cotton, the most popular crops are maize, rice, groundnuts and cassava. Average total annual household expenditure on hired labor, fertilizer, pesticides, animal rental for plowing, crop transport and other miscellaneous inputs was 150,156 Tanzanian shillings, which is about \$107 US at 2008-2009 exchange rates.

In this paper we make use of only a subset of the available subjective distributions and production data. We opt for a sparse specification of the production function because we are concerned primarily with development of a procedure that incorporates the subjective plot-level distributions and farmer-level price distributions into the estimation algorithm. We use the following plot-level variables: acreage, labor inputs in the first and second half of the cultivation period, final output quantity, and a sequence of subjective distributions over final output from different points in the cultivation year. We also use a sequence of price expectations observed at the farmer level (and therefore, for our purposes, the plot level), and median village labor prices. Labor inputs are aggregated across various survey rounds. The average length of time between the first and last application of non-planting, non-harvest labor on the sample plots is 21 weeks. Therefore, the variable l_1 in the model below captures the average number of total person-days per week over the first 12 weeks of post-planting cultivation; l_2 is the weekly average person-days over the next 9 weeks.⁶ We make no attempt to control for

⁶The number of weeks used to determine l_2 varies slightly for a small number of plots

Table 5.1: Sample summary statistics

	Mean	sd	Min	Max
Household size (people)	8.33	3.90	2	23
Dependency ratio*	1.31	0.85	0	5.5
Head age	46.85	14.69	20	100
Head is male (%)	85.0	-	-	-
Years of education (HH head)	4.19	3.46	0	11
Radios	0.83	0.71	0	4
Bicycles	1.19	1.00	0	10
Dairy cattle	1.33	2.84	0	20
Non-dairy cattle	3.87	7.89	0	60
Goats	5.27	8.05	0	50
Sheep	1.67	3.74	0	30
Total acres	9.67	11.03	1	82
Number of plots	2.71	1.17	1	7
Number of crops grown	3.45	1.26	1	8
Labor expenditure (TSH)	78,248	139,485	0	1,020,000
Fertilizer expenditure (TSH)	21,149	81,359	0	715,000
Animal labor expenditure (TSH)	33,497	92,724	0	750,000
Transport expenditure (TSH)	10,333	20,049	0	144,000
Other cultivation expenditure (TSH)	6,929	15,817	0	100,000
Total cultivation expenditure (TSH)	150,156	254,863	0	1,514,700

Notes: author's calculation from survey data; cultivation data refers to 2008-2009 cultivation of all crops; 1 USD \approx 1,400 TSH; *Dependency ratio is number of persons aged < 15 or aged > 65 divided by number aged between 15 and 65.

Table 5.2: Plot-level variables used in the estimation (N=212)

Variable	Mean	s.d.
Acres	2.04	1.62
Labor 1 (avg weekly person-days)	3.29	0.29
Labor 2 (avg weekly person-days)	2.96	0.47
Expected price in period 1 (TSH / <i>mafurushi</i>)	48,247	5,528.3
Expected price in period 2 (TSH / <i>mafurushi</i>)	51,021	5,311.1
Output (1 <i>mafurushi</i> = 90 kg)	5.78	7.34

farmer-level variation, other than through the expected price variable that is common to plots belonging to a single farmer.

We dropped any plots for which any of the labor, output or subjective distributions information data are missing or coded as “Don’t know”. We also dropped 4 plots at the estimation stage whose input, output and distributional data were in such conflict with the model that they made no contribution to the log likelihood for any reasonable combination of parameters.⁷ We ended up with a total of 212 cotton plots with which to estimate the model. Table 5.2 reports summary statistics for the variables included in the estimation stage.

that were harvested very early.

⁷Data for at least one of these plots was given by different respondents at different times, and the plot-level output distributions vary in a way that is completely inconsistent with a model in which information is revealed gradually to a profit-maximizing farmer.

5.3 Stochastic Production Model

Farmers face substantial production uncertainty caused by agro-climatic factors that are beyond their control, such as rainfall levels, temperature, pest pressure, destruction by birds and wild animals, and unobserved components of soil quality. Other shocks, such as illness, may further disrupt otherwise optimal production plans. In this section we incorporate the role of exogenous shocks and the gradual resolution of temporal uncertainty into the farmer's decision problem. We develop a stochastic production model that deals explicitly with the sequential nature of farmers' input allocation decisions, and with the gradual revelation of information about the state of nature.

In order to focus on the role of subjective probability distributions, we maintain numerous simplifying assumptions. We treat household consumption and production decisions as fully separable, so that farmers act as risk-neutral expected profit maximizers over their final output from cotton cultivation. We assume that input prices are known and fixed, and that interest-free credit markets and labor reciprocity norms are sufficiently well-developed to satisfy labor demand in every period. Furthermore, although there are inputs other than labor (notably, pesticides) that play an important role in cotton production, in this paper we model crop growth as a function only of labor and stochastic shocks. This allows us to use the same crop growth specification as that in Fafchamps (1993).

We divide the cultivation period into 4 stages: Planting, Cultivation 1,

Cultivation 2, and Harvest. We number these as periods 0 – 3, respectively. Harvest-period labor is assumed to be directly proportional to output, therefore we ignore the final period. Likewise, while planting labor is potentially informative for crop growth, we take acreage as exogenous and ignore the role of labor input decisions during the planting period. We do, however, allow for a stochastic shock to affect crop growth after planting but prior to the application of cultivation labor. This shock corresponds to negative weather and pest shocks related to the plants’ “take” and the possible need for re-seeding.

Final output on plot i , denoted y_i , is a function of acreage A_i , labor in periods 1 and 2, and stochastic shocks in periods 0, 1 and 2. Following Fafchamps (1993), it will be convenient to write the production function as a sequence of functions, in which the crop state in period t is determined by the lagged state of crop growth, period t labor inputs, and the contemporaneous stochastic term:

$$\begin{aligned}
 y_{i0} &= A_i e^{\theta_{i0}} \\
 y_{i1} &= h_1(y_{i0}, l_{i1}) e^{\theta_{i1}} \\
 y_i &= h_2(y_{i1}, l_{i2}) e^{\theta_{i2}}
 \end{aligned} \tag{5.1}$$

where $\theta_{it} \sim g_{it}(\theta_{it})$ for $t = 0, 1, 2$

Throughout the paper, we use $\psi_{it}(y_i)$ to denote the subjective distribution over final output y_i gathered in cultivation stage t . This is in contrast

to $g_{it}(\theta_{it})$, which are the densities of the stochastic contributions to output. The parameters of the crop growth functions h_t , $t = \{1, 2\}$, are allowed to vary across periods, to capture the changing role of labor across the cultivation season. We use constant returns CES production functions in order to maintain flexibility in the sign and size of elasticities of substitution without substantially increasing the number of parameters to be estimated:

$$h_1(y_0, l_1 | \alpha, \gamma) = [\alpha y_0^\gamma + (1 - \alpha) l_1^\gamma]^\frac{1}{\gamma} \quad (5.2)$$

$$h_2(y_1, l_2 | B, \beta, \delta) = B[\beta y_1^\delta + (1 - \beta) l_2^\delta]^\frac{1}{\delta} \quad (5.3)$$

There are a total of 5 production parameters: the share parameters α and β which lie in the interval $(0, 1)$, the transformed elasticity parameters γ and δ , and the global scale parameter B . While the problem is also dependent on the period- and plot-specific distributions of the stochastic terms, i.e., the parameters of g_{it} , we do not parameterize these densities because close approximations of them are observed and therefore do not need to be estimated. We discuss this issue in detail in the following section.

The farmer's objective in period 1 is to maximize expected plot-level profits conditional on the observed level of θ_{i0} , on the subjective distributions over θ_{i1} and θ_{i2} , and on his expectations about his own future labor input

l_{i2} :

$$\max_{l_{i1}} E[q_c] E_{\theta_{i1}\theta_{i2}} \left[B \left\{ \beta \left([\alpha (A_i e^{\theta_{i0}})^\gamma + (1-\alpha) l_1^\gamma]^\frac{1}{\gamma} e^{\theta_{i1}} \right)^\delta + (1-\beta) l_{i2}^{*\delta} \right\}^\frac{1}{\gamma} e^{\theta_{i2}} \right] - q_l l_{i1} \quad (5.4)$$

where the expectation $E_{\theta_{i1}\theta_{i2}}$ is taken over $g_{i1}(\theta_{i1})$ and $g_{i2}(\theta_{i2})$, the future choice of l_{i2} is replaced with its optimal policy function l_{i2}^* , and the input price q_l is known and fixed. A similar expression can be written for labor in period 2.

Clearly, the timing of information revelation and labor decisions plays an important role in the model. Therefore, before going further we discuss the sequence of events that characterize the farmer's problem. We will derive the model's first-order conditions in the section covering the identification of the error densities $g_{it}(\theta_{it})$. In the rest of the paper we will usually drop the plot subscript i , for notational simplicity, but we nevertheless maintain the plot- and period-level specificity of the distributions of the stochastic contributions to output.

5.3.1 Timing

The timing of input allocation decisions and information revelation is critical to the model. Therefore, before moving on to consider the farmer's problem and the identification of the subjective distributions over the sequence of shocks, we make the sequence of events explicit.

The first event is the determination of plot acreage A . In this paper we

treat acreage as exogenous, so period 0 of the farmer's decision problem begins after A is chosen.⁸ Prior to the application of weeding and trimming labor, i.e. prior to the choice of l_1 , stage 0 yield distribution $\psi_0(y)$ is reported by the farmer to the researcher. At this point the farmer incorporates his rational beliefs about his own future actions, given by optimal policy functions l_1^* and l_2^* , as well as his beliefs about the densities of future shocks, $g_1(\theta_1)$ and $g_2(\theta_2)$, into the distribution over final output $\psi_0(y)$ that he communicates to the researcher.

The first shock θ_0 is then realized, and period 1 begins. The farmer observes the realization of θ_0 , and chooses l_1 . He makes this choice in accordance with density functions $g_1(\theta_1)$ and $g_2(\theta_2)$, and with the optimal policy function l_2^* that determines his future labor input decision. Output distribution $\psi_1(y)$ is observed by the researcher simultaneously with l_1 , therefore it includes the actual choice l_1 , rather than the optimal policy function l_1^* .

The period 1 shock θ_1 is then realized, and period 2 begins. The farmer observes the realized θ_1 , and chooses l_2 in accordance with his beliefs $g_2(\theta_2)$. At this point he also reveals $\psi_2(y)$ to the researcher. Therefore, subjective distribution $\psi_2(y)$ includes the realized values of θ_0, θ_1, l_1 and l_2 . The only stochastic variation remaining is that which is revealed at the end of period 2, when θ_2 is realized. This last shock captures both the possibilities of

⁸We do observe yield distributions that pre-date the acreage decision, but they have yet to be incorporated into the model. The early season yield distributions were gathered using a different set of bin-intervals than the later distributions. Comparability between the yield distributions from before and after the acreage decision will be improved once we relax the stepwise uniform assumption.

continued growth or degradation in the final weeks before harvesting, and the residual uncertainty over actual output volume once harvesting is complete.

It should be emphasized that this schedule of information revelation, data collection and labor allocation is fully observed in the data. Labor inputs that are aggregated across multiple rounds of data collection are organized to match the elicitation of subjective yield distributions in a manner that corresponds to the above sequence.

5.4 Estimation Procedure

In this section we describe the procedure used to estimate the five production parameters of the stochastic production model. Because this paper's key innovation involves the use of a sequence of subjective plot-level output distributions to identify the distributions of stochastic contributions to output, thereby avoiding the restrictive assumption of a common error variance across plots and periods, we first discuss in detail the identification of the distributions of $(\theta_0, \theta_1, \theta_2)$.

5.4.1 Identification of error density functions

If we assume that farmers have rational expectations about their own future input choices, the density functions $g_0(\theta_0)$, $g_1(\theta_1)$ and $g_2(\theta_2)$ are identified by the sequence of observed, subjective output distributions. The key is that period t and period $t+1$ distributions over final output differ in only two ways:

the former incorporates θ_t as a random variable and l_{t+1} as an optimal policy function, while the latter treats these as an observed realization of a shock and a known labor allocation decision, respectively. Working backwards from $\psi_2(y)$, we will show that with a CES production function specification, the sequence of observed yield distributions fully reveals the error distributions.

First, some additional assumptions about the sequence of shocks are needed in order to identify $g_0(\theta_0)$, $g_1(\theta_1)$ and $g_2(\theta_2)$ from $\psi_0(y)$, $\psi_1(y)$ and $\psi_2(y)$. These are as follows:

1. The plot-specific error densities $g_0(\theta_0)$, $g_1(\theta_1)$ and $g_2(\theta_2)$ are known throughout the cultivation season and are independent of inputs l_1 and l_2
2. The sequence of shocks $(\theta_0, \theta_1, \theta_2)$ are mutually independent (though not necessarily identically distributed)
3. $E[e^{\theta_t}] = 1 \quad t = 0, 1, 2$

For all t , assumptions 1 and 2 rule out the possibility of within-season learning about the conditional or the unconditional distribution of θ_t (which are identical). This does not preclude the possibility of between-season learning. Presumably, farmers' beliefs about the distributions of the shocks are developed over a period of years from their own experiences and the experiences of their neighbors. However, within season the farmer learns the realization of θ_t , but that tells him nothing about the distribution of θ_{t+1} .

Likewise, because the unconditional error distribution is the same as the error distribution conditioned on inputs, the farmer cannot take action to change the variance of future shocks.⁹ Assumption 3 is a location normalization; it will turn out to be more convenient to use this normalization than to make the usual assumption $E[\theta_t] = 0$, for $t = 0, 1, 2$.

Lastly, we also assume that the subjective output distributions that are reported by farmers to the researcher, $\psi_t(y)$, are the true output distributions. This is analogous to a standard rational expectations assumption, with the critical difference that here we need only assume that the farmer knows the error distributions and does his best to communicate them to the researcher. We do not need to assume that the shock distributions are identical across periods and plots, nor that they vary systematically in accordance with a small number of parameters, both of which are standard in dynamic resource allocation models. In fact, we show below that there is substantial heterogeneity in error distributions, across plots and across time.

To derive the error densities from the output densities, consider first the observed plot-specific period 2 output distribution $\psi_2(y)$. When ψ_2 is elicited, all inputs have been allocated and all shocks other than θ_2 have been realized and observed. Let Ω_2 denote the information set associated with this period.

⁹The interaction between input allocations and subjective, conditional error distributions is an interesting avenue for future research. It has long been known that many of the functional forms imposed on agricultural production data impose overly restrictive assumptions on the relationship between inputs and the variance of output (Just and Pope, 1978).

Ω_2 includes $\psi_2(y)$. Using the production function given in (5.1), we can write:

$$\begin{aligned} E[y|\Omega_2] &= E[Bh_2(A, l_1, l_2, \theta_0, \theta_1; \nu)e^{\theta_2}|\Omega_2] = E[Bh_2(A, l_1, l_2, \theta_0, \theta_1; \nu)] \cdot E[e^{\theta_2}] \\ &= Bh_2(A, l_1, l_2, \theta_0, \theta_1; \nu) \end{aligned} \quad (5.5)$$

where $\nu = (\alpha, \beta, \gamma, \delta)$. In the above we make use of the independence of θ_2 and the normalization $E[e^{\theta_2}] = 1$.

Because the deterministic portion of production is fully known at the time $\psi_2(y)$ is elicited, the expected value of final output y is equivalent to the expectation of y conditional on the observed output density $\psi_2(y)$. The distributional information in $\psi_2(y)$, then, reveals information about the density of the only remaining stochastic contribution to output, θ_2 . Recall that $\psi_2(y)$ takes the form of a measure vector p_2 of length N associated with N connected and non-overlapping intervals on the support of y , and that we assume the density inside any interval follows a uniform distribution. For the interval j , defined by boundaries $[a_j, b_j]$, $j = 1 \dots N$, we observe $p_{2j} = \Pr[a_j \leq y \cap y \leq b_j]$. Incorporating (5.5), we can make the following substitution:

$$\begin{aligned} p_{2j} &= \Pr \left[a_j \leq y \cap y \leq b_j \right] \quad \text{in period 2} \\ &= \Pr \left[a_j \leq E[y|\Omega_2]e^{\theta_2} \cap E[y|\Omega_2]e^{\theta_2} \leq b_j \right] \\ &= \Pr \left[\ln \left(\frac{a_j}{E[y|\Omega_2]} \right) \leq \theta_2 \cap \theta_2 \leq \ln \left(\frac{b_j}{E[y|\Omega_2]} \right) \right] \end{aligned} \quad (5.6)$$

Thus, through the straightforward transformation of the interval boundaries given by (5.6), $\psi_2(y)$ becomes $g_2(\theta_2)$, and the associated measure vector is p_2 .

We now consider the identification of $g_1(\theta_1)$. At the time output distribution $\psi_1(y)$ was collected, the realization of θ_0 and the known choice l_1 were already determined. Stochastic contributions to $\psi_1(y)$ are from the farmer's subjective distribution over the two remaining shocks, θ_1 and θ_2 , as well as the optimal policy function for future labor, l_2^* , which depends on θ_1 . We already have a measure of the θ_2 density $g_2(\theta_2)$. This subjective distribution is time invariant, therefore it is part of Ω_1 , the period 1 information set. To derive an expression for l_2^* we return momentarily to the farmer's choice problem in period 2. The farmer is an expected profit maximizer, therefore his period 2 problem is given by:¹⁰

$$\max_{l_2} E[q_c] E_{\theta_2} \left[B \left\{ \beta \left([\alpha (A e^{\theta_0})^\gamma + (1-\alpha) l_1^\gamma]^\frac{1}{\gamma} e^{\theta_1} \right)^\delta + (1-\beta) l_2^\delta \right\}^\frac{1}{\gamma} e^{\theta_2} \right] - q_l l_2 \quad (5.7)$$

where q_c is the price of cotton at the time of sale, q_l is the known price of labor in period 2, and the second expectation is taken only over θ_2 , because this problem is solved after θ_1 has been realized. In (5.7) we rely on the observation, discussed in Section 5.2 that Tanzanian cotton farmers are pure price-takers in the output market, so that plot-level output and prices are

¹⁰Recall that we have assumed a fully separable model with interest-free credit markets sufficient to cover the cost of any desired labor inputs. Therefore only the contemporaneous cost of labor enters the expected profit equation.

independent. The term $E[q_c]$ is observed, because the survey data includes regular measures of subjective output price distributions. Likewise, q_l is observed directly in the data set. After taking the first order condition of (5.7) and rearranging, the l_2 optimal policy function is given by:

$$l_2^*(A, l_1, E[q_c], q_l, \theta_0, \theta_1, \eta, B) = \frac{\beta(\alpha(Ae^{\theta_0})^\gamma + (1-\alpha)l_1^\gamma)^{\frac{1}{\gamma}} e^{\theta_1}}{\left(\frac{q_l}{BE[q_c](1-\beta)}\right)^{\frac{\delta}{1-\delta}} - (1-\beta)} \quad (5.8)$$

where $\eta = (\alpha, \beta, \gamma, \delta)$. Using (5.8), we return to period 1 and write output as a function of l_1 and the shocks:

$$\begin{aligned} y &= B \left(\beta \{ (\alpha(Ae^{\theta_0})^\gamma + (1-\alpha)l_1^\gamma)^{\frac{1}{\gamma}} \}^\delta e^{\delta\theta_1} + \frac{(1-\beta)\beta \{ \alpha(Ae^{\theta_0})^\gamma + (1-\alpha)l_1^\gamma \}^{\frac{\delta}{\gamma}} e^{\delta\theta_1}}{\left(\frac{q_l}{BE[q_c](1-\beta)}\right)^{\frac{\delta}{1-\delta}} - (1-\beta)} \right)^{\frac{1}{\delta}} e^{\theta_2} \\ &= B \left(\beta \{ (\alpha(Ae^{\theta_0})^\gamma + (1-\alpha)l_1^\gamma)^{\frac{1}{\gamma}} \}^\delta + \frac{(1-\beta)\beta \{ \alpha(Ae^{\theta_0})^\gamma + (1-\alpha)l_1^\gamma \}^{\frac{\delta}{\gamma}}}{\left(\frac{q_l}{BE[q_c](1-\beta)}\right)^{\frac{\delta}{1-\delta}} - (1-\beta)} \right)^{\frac{1}{\delta}} e^{\theta_1} e^{\theta_2} \\ &= H_1(B, \alpha, \beta, \gamma, \delta; A, l_1, q_l, E[q_c], \theta_0) e^{\theta_1} e^{\theta_2} \end{aligned} \quad (5.9)$$

Substituting the l_2^* policy function into the period 1 output equation allows us to factor out e^{θ_1} , and express output as the product of H_1 , which is a function of parameters, pre-determined variables and the choice of l_1 , and shock terms that enter proportionally. This is a general feature of any production function the log of which is additive in the shocks: substitution of the optimal policy functions from future decision periods into the expected profit equation allows us to express output as the product of its deterministic

and stochastic components. Because the shocks θ_1 and θ_2 are independent of each other and H_1 is non-stochastic, we can write an expression for expected output that is analogous to (5.5):

$$\begin{aligned}
 E[y|\Omega_1] &= E[H_1(B, \alpha, \beta, \gamma, \delta; A, l_1, q_l, E[q_c], \theta_0) e^{\theta_1} e^{\theta_2} | \Omega_1] \\
 &= E[H_1(B, \alpha, \beta, \gamma, \delta; A, l_1, q_l, E[q_c], \theta_0) | \Omega_1] \cdot E[e^{\theta_1}] \cdot E[e^{\theta_2}] \\
 &= H_1(B, \alpha, \beta, \gamma, \delta; A, l_1, q_l, E[q_c], \theta_0)
 \end{aligned} \tag{5.10}$$

Once again, the deterministic component of output is observed as the expectation of y given by $\psi_1(y)$. The distributional information in $\psi_1(y)$ that describes the subjective distribution of y around its mean is based entirely on the multiplicative effects of e^{θ_1} and e^{θ_2} . Using the observed period 1 measure vector p_1 , we can write the following expression for the probability of final output y taking on a value in interval j :

$$\begin{aligned}
 p_{1j} &= \Pr \left[a_j \leq y \cap y \leq b_j \right] \quad \text{in period 1} \\
 &= \Pr \left[a_j \leq E[y|\Omega_1] e^{\theta_2} e^{\theta_1} \cap E[y|\Omega_1] e^{\theta_1} e^{\theta_2} \leq b_j \right] \\
 &= \Pr \left[E[y|\Omega_1] e^{\theta_1} e^{\theta_2} \leq b_j \right] - \Pr \left[a_j \leq E[y|\Omega_1] e^{\theta_2} e^{\theta_1} \right] \\
 &= \Pr \left[\theta_1 \leq \ln \left(\frac{b_j}{E[y|\Omega_1]} \right) - \theta_2 \right] - \Pr \left[\ln \left(\frac{a_j}{E[y|\Omega_1]} \right) - \theta_2 \leq \theta_1 \right]
 \end{aligned} \tag{5.11}$$

The expression in (5.11) suggests a method for numerically estimating the distribution of θ_1 . Imagine momentarily that we knew $g_2(\theta_2)$ and $g_1(\theta_1)$,

but did not know the distribution of y . Because y is a function of two independent random variables, a consistent estimator of the distribution of y conditional on $g_2(\theta_2)$ and $g_1(\theta_1)$ can be derived by repeatedly sampling from g_1 and g_2 , calculating the y associated with each sampled (θ_1, θ_2) pair, and estimating the distribution of the calculated y 's. In our case, we already have expressions for the distribution of y and θ_2 , $\psi_1(y)$ and $g_2(\theta_2)$, respectively, but it is the distribution of θ_1 that is unknown. We do, however, observe $E[y|\Omega_1]$ as the expectation of y conditional on $\psi_1(y)$. Because we know that y is given by (5.10) and we assume that the shocks are independent, we can approximate the distribution of θ_1 by drawing M pairs (y_m, θ_{2m}) from $\psi_1(y)$ and $g_2(\theta_2)$, and calculating the following:

$$\Pr[\theta_1 < \Theta_1] = \frac{1}{M} \sum_{m=1}^M \mathbb{I} \left[\ln \left(\frac{y_m}{E[y|\Omega_1]} \right) - \theta_{2m} \leq \Theta_1 \right] \quad (5.12)$$

where \mathbb{I} is the indicator function, for any value of Θ_1 on the real line. From the empirical CDF given by (5.12) we can easily reconstruct an approximation of $g_1(\theta_1)$ that follows the form of (5.11). For each plot, we set $M = 20,000$ and empirically approximate $g_1(\theta_1)$ in just this fashion. To maintain consistency with the stepwise uniform functional form of $\psi_1(y)$ and $g_2(\theta_2)$, we approximate $g_1(\theta_1)$ as stepwise uniform over 20 intervals, each corresponding to a 5% quantile.¹¹ Because we select the interval boundaries to ensure equal probability in each interval, the width of the intervals that

¹¹In order to bound the upper and lower intervals in a reasonable way, we throw out the upper and lower 0.005% tails of the M estimates of θ_1 before constructing interval boundaries.

define $g_1(\theta_1)$ is not standardized.

A notable feature of this estimation method is that the estimate of $g_1(\theta_1)$ does not depend on production parameters. This is a direct result of the CES functional form, which allowed us to factor output into its deterministic and stochastic components. Independence from model parameters is not a necessary condition for consistency of (5.12), but it does save substantial computer time, since g_1 can be estimated once for each plot and then stored for use in each iteration of the search algorithm. It appears to be the case that monotonicity of y in θ_1 is a sufficient condition for consistency of (5.12), regardless of whether the expression inside \mathbb{I} in (5.12) depends on model parameters.

We use the same procedure to derive the density function for θ_0 , $g_0(\theta_0)$. After substituting l_2^* into the period 1 problem and solving for optimal policy function l_1^* , we can express expected output as follows:

$$E[y|\Omega_0] = H_0(B, \alpha, \beta, \gamma, \delta; A, q_l, E[q_c])e^{\theta_0}e^{\theta_1}e^{\theta_2} \quad (5.13)$$

We then sample repeatedly from $\psi_0(y)$ and our estimates $g_1(\theta_1)$ and $g_2(\theta_2)$, and construct an empirical approximation of $g_0(\theta_0)$ following the same method as in (5.12).

5.4.2 Estimation Algorithm

Using the estimated density functions for stochastic shocks $(\theta_0, \theta_1, \theta_2)$, the joint probability of observing the sample conditional on parameter vector $\eta = (B, \alpha, \beta, \gamma, \delta)$ is given by:

$$\mathbb{L}(A, l_1, l_2, Y|\eta) \prod_{i=1}^P g_{i0}(\theta_{i0})g_{i1}(\theta_{i1})g_{i2}(\theta_{i2}) \quad (5.14)$$

where P is the total number of plots in the sample and the density functions are indexed to emphasize their plot-specificity. Equation (5.14) is analogous to a concentrated likelihood function in a standard model without subjective distributional data, because error parameters do not enter directly.

In order to evaluate this expression, we first need to explain how estimates of realized $(\theta_0, \theta_1, \theta_2)$ depend on the model parameters. In the case of θ_0 and θ_1 , the estimated realized shocks are calculated as those that satisfy l_1^* and l_2^* . That is, for a given parameter vector we find the value of the realized shock that must have been observed by the farmer in order to make his observed labor allocation decision correspond to that given by his optimal policy rule. Because l_2^* depends on θ_0 , we first find θ_0 using the optimal policy function

for l_1 :

$$l_1^*(B, \alpha, \beta, \gamma, \delta, A, \theta_0, E[q_c], q_l) = \frac{\alpha^{\frac{1}{\gamma}} A e^{\theta_0}}{\left[\left(\frac{q_l}{1-\alpha} \right)^{\frac{\gamma}{1-\gamma}} \left\{ \frac{BE[q_c] \beta^{\frac{1}{\delta}} \Sigma}{\Sigma - 1 + \beta} \right\}^{\frac{\gamma}{\gamma-1}} - 1 + \alpha \right]^{\frac{1}{\gamma}}} \quad (5.15)$$

$$\text{where } \Sigma = \left(\frac{q_l}{BE[q_c](1-\beta)} \right)^{\frac{\delta}{1-\delta}}$$

For any value of the parameter vector, the unique value of θ_0 that satisfies (5.15) can be quickly calculated. In practice, θ_0 takes on complex values over substantial portions of the parameter space, and an extensive grid search is needed to identify reasonable starting values for the parameters prior to implementing an iterative search procedure. Once we have an estimate of θ_0 , we can quickly recover an estimate of θ_1 from (5.8). Lastly, θ_2 is estimated as $\theta_2 = \ln\left(\frac{Y}{H_2(\cdot)}\right)$, where $H_2(\cdot)$ is the right hand side of (5.5).

Using these estimates of the shocks, we search for the value of the parameter vector that maximizes the log of (5.14).¹² The density functions over the stochastic components of output are observed and invariant to parameter values, therefore calculation of the log likelihood for successive guesses of pro-

¹²Is is noteworthy that while contributions to the log likelihood are independent across plots, which is standard, the contribution in each period is also independent of future shocks on the same plot. This suggests that early stage data from plots for which input data is incomplete in later rounds, due to missing values or missed interviews, could in principle be added to the log likelihood. Of course, this requires that subjective output distributions are never missing even when input data are, because the error densities are only identified via backwards induction from the final period.

duction parameters $(B, \alpha, \beta, \gamma, \delta)$ is very rapid. However, because we have imposed a stepwise uniform distributional assumption on the density functions, the gradient vector consists only of 0 and ∞ values, making Hessian-based search methods infeasible. We therefore make use of an uphill simplex method (also called the amoeba) similar to the Nelder-Mead algorithm, which does not require calculation of the gradient or Hessian of the log likelihood (Press *et al.*, 2007).

5.4.3 Standard errors

In principle, the covariance matrix of the estimated parameter vector can be estimated in the standard fashion, i.e., with the inverse of the information matrix evaluated at the estimated parameter value. However, it is an open question whether the subjective nature of the error distributions changes the asymptotic distribution of the ML estimate. Regardless, in the current context the standard estimate of the asymptotic variance of the estimator is not calculable, because our distributional assumptions about the g_t functions do not permit analytical computation of the information matrix. We therefore focus on estimation of the parameter vector itself, and leave its asymptotic distribution for future work.

5.5 Results

We focus on two sets of results: the error distributions and the estimated parameter vector. While the former are not in fact results, because they are observed in the data and are “estimated” by transforming the subjective output distributions, they demonstrate the degree of heterogeneity in beliefs that would be lost if a common-error assumption were imposed on these data. Table 5.3 shows the average and the standard deviation of the upper and lower bounds and expected value of the estimated densities of θ_0 , θ_1 and θ_2 . Note that non-zero average expected values are expected, because we normalized these densities so that the expectation of the exponential is unity. We are particularly interested in the standard deviation column of Table 5.3. There is clearly a substantial degree of within-period variation in both the intervals of positive support and the expected values of the plot-specific distributions. This suggests that farmers do in fact perceive variation in the stochastic contribution to production on plots of different characteristics and in accordance with the farmers’ plot-specific cultivation plans. Rational expectations models that suppress this variation will, in general, arrive at biased estimates of the parameter vector.

In Table 5.4 we see the estimated coefficient vector and the value of the log-likelihood function at its highest point. Only two decimal digits were estimated, because a relatively slack convergence tolerance was used to generate these results, starting from numerous initial points identified by grid

Table 5.3: $g_t(\theta_t)$ summary statistics

Variable	Mean	s.d.
θ_0 lower bound	-2.95	1.86
θ_0 upper bound	2.43	1.70
$E[\theta_0]$	-0.14	0.61
θ_1 lower bound	-2.49	1.84
θ_1 upper bound	2.01	1.48
$E[\theta_{01}]$	-0.01	0.58
θ_2 lower bound	-4.19	1.4807*
θ_2 upper bound	3.19	1.4807*
$E[\theta_2]$	-2.35	1.17
N	212	212

*SD of θ_2 upper and lower bounds is constant by construction, because both reflect variation in acreage

Table 5.4: Estimated parameter vector

Parameter	Value
B	5.51
α	0.21
β	0.01
γ	-1.84
δ	-4.87
Log likelihood	-3,044.20

search. The most notable results are those for B and β : the scale parameter plays a very significant role in matching realized output to the approximate center of the output distributions $\psi_2(y)$. The very low value of β suggests that labor inputs in the second half of the season have an outsized effect on final output. It is notable that Fafchamps (1993) found a similarly near-degenerate share parameter value, though in that paper early stage crop growth figured prominently and early stage labor had a near-zero share.

While we can be confident that the estimated vector is a close approximation of the global maximum associated with the model estimated in this paper, the log-likelihood value indicates that the model is a relatively poor fit of the data. This is not surprising, as we have ignored relevant input data (pesticides) and imposed a risk-neutrality assumption that is surely inconsistent with the choice behavior of many farmers. These shortcomings, however, can be remedied in future versions of this paper. It will also be interesting to estimate this model again, ignoring the observed distributions and imposing a standard rational expectations model. It is likely that if we treated the θ_t variances as free parameters that could adjust to fit the structural parameters, we could improve the fit of a model that is clearly inconsistent with the observed subjective distributions.

5.6 Conclusion

In this paper we developed a technique for identifying subjective plot-level distributions of stochastic shocks to agricultural output from a sequence of plot-level yield distributions, and applied the technique to estimate a dynamic agricultural production model from a high frequency data set of Tanzanian cotton farmers. The distributional data indicate a substantial degree of heterogeneity in perceived error distributions, across both plots and periods. This suggests that well-specified models that utilize measured subjective distributions will, in general, arrive at parameter estimates that are

uncontaminated by the largely untestable assumptions that have been the workhorse for this type of model for decades.

One of the important take-away lessons of the paper is that the aggregated contributions of shocks that are subject to considerable inter-plot variation and that are difficult to quantify, e.g. pest pressure, can still be identified by proper transformation of a sequence of subjective output distributions. While we cannot disaggregate the plot-period-level shocks into the components that are due to rainfall uncertainty, pest uncertainty, soil quality, etc., we can derive an estimate of a composite error directly from the elicited subjective output distributions. This has important implications for the design of future efforts to gather subjective distributions data. Arbitrary quantification of shocks that are perceived qualitatively or categorically may not be necessary if distributions over a related, well-conceived continuous variable can be gathered.

5.7 Appendix

Included with this chapter are the phone survey instrument used to gather high frequency input, output distribution and price distribution data. We also include the visual aids used to elicit output and price distributions, laminated copies of which were left with each respondent.

Figure 5.1: Phone survey instrument, Page 1

Figure 5.2: Phone survey instrument, Page 2

PLOT #					
7. From whom did you buy these pesticides? <div>TCB.....1 Private seller.....2 Other (specify).....3</div>	8. Have you sprayed any more pesticides on this plot, during the last 6 weeks? YES...1 NO...2 >> 22	9. In what week and month did you do the SECOND spray of pesticides on this plot (during the last 6 weeks)? WEEK MONTH	10. How many liters did you spray the SECOND time? NUMBER OF LITERS	11. Was the spray oil-based or water-based? <div>Oil.....1 Water...2</div>	12. How much did you pay in total for the pesticides used IN THIS SPRAY, ON THIS PLOT? TSH

Figure 5.3: Phone survey instrument, Page 3

Figure 5.4: Phone survey instrument, Page 4

Figure 5.5: Phone survey instrument, Page 5

Figure 5.6: Phone survey instrument, Page 6

[illegible]

Figure 5.8: Phone survey instrument, Page 8

Figure 5.9: Phone survey instrument, Page 9

[illegible]

PRICE EXPECTATIONS						
45. I'd like you to think about the cotton price that you expect to receive later this year (not the TCB minimum, since they could be different). You will need fourteen BEANS OR SEEDS. Please turn to Page #10 of the sheets that we left you. Think about the various factors that might affect the price you receive next year, and the possibilities of receiving a high price, a middle price or a low price. Divide the 14 seeds among the boxes according to the price that you expect to receive when you sell your cotton next year.						
ENTER "0" IF NO SEEDS PLACED IN A BOX						
CHECK THAT THE TOTAL NUMBER OF SEEDS IS 14						
280-340	340-400	400-460	460-520	520-580	580-640	640-700+

Figure 5.11: Phone survey instrument, Page 11

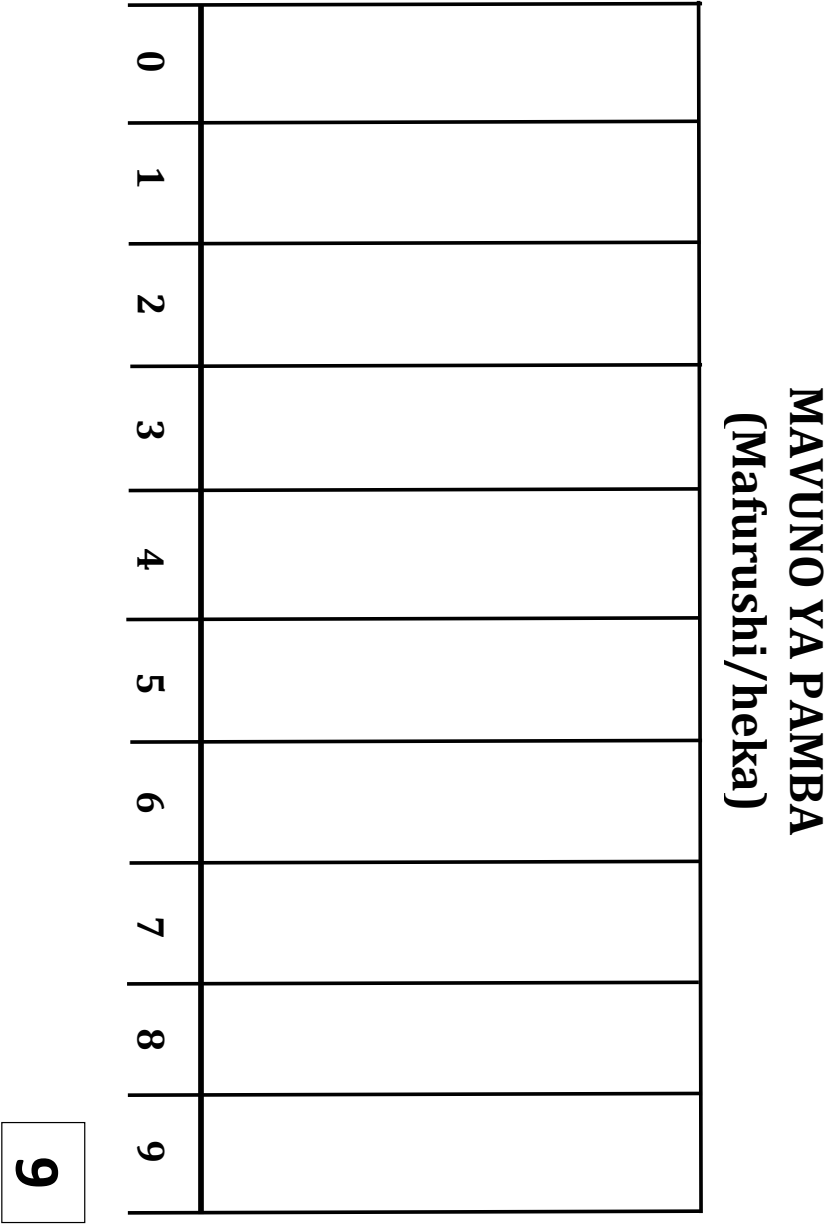


Figure 5.12: Visual aid used to elicit yield distributions

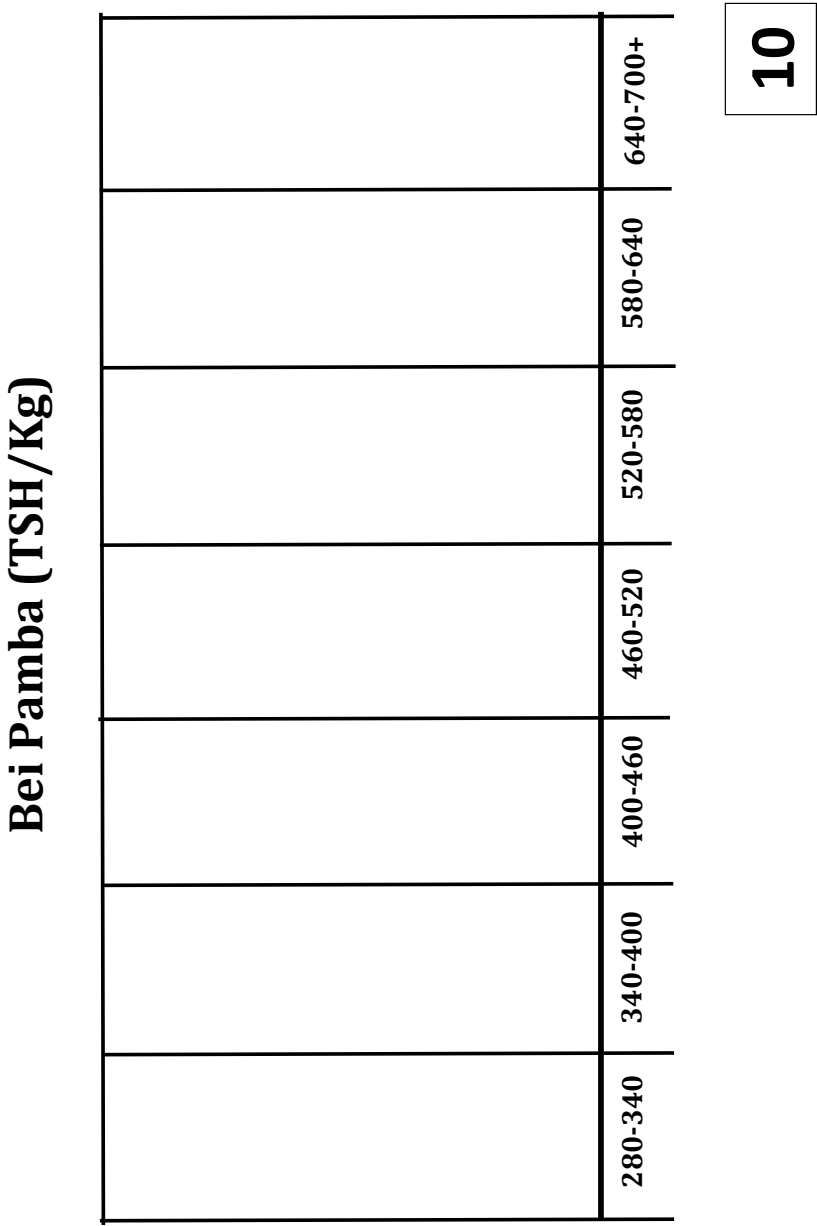


Figure 5.13: Visual aid used to elicit price distributions

REFERENCES

- Attanasio, O and K Kaufmann (2009). "Educational Choices, Subjective Expectations and Credit Constraints", Working paper.
- Boissiere, M, JB Knight and RH Sabot (1985). "Earnings, Schooling, Ability and Cognitive Skills", *American Economic Review* 75(5): 1016-1030.
- Bradu, D and Y Mundlak (1970). "Estimation in Lognormal Linear Models", *Journal of the American Statistical Association* 65(329): 198-211.
- Camacho and Conover (2011). "The Impact of Receiving Price and Climate Information in the Agricultural Sector", IDB Working Paper Series No. IDB-WP-220.
- Card, D (1999). "The Causal Effect of Education on Earnings", in O Ashenfelter and D Card, eds., *Handbook of Labor Economics, Volume 4*. Elsevier.
- Carneiro, P, J Heckman and E Vytlačil (2010). "Estimating Marginal Returns to Education", NBER Working Papers #16474.
- Chamberlain, G (1980). "Analysis of covariance with qualitative data", *Review of Economic Studies* 47: 225-238.
- Cole, S and S Hunt (2010). "Information, Expectations, and Agricultural Investments: Evidence from a Field Experiment in India." Working paper.
- Delavande, A, X Gine and D McKenzie (2010). "Measuring Subjective Expectations in Developing Countries: A Critical Review and New Evidence." *Journal of Development Economics* 94(2): 151-163.
- Delavande, A, X Gine and D McKenzie (2011). "Eliciting Probabilistic Expectations with Visual Aids in Developing Countries: How sensitive are answers to variations in elicitation design?" *Journal of Applied Econometrics* 26(3): 479-497.
- Delavande, A and H Kohler (2008a). "Subjective Expectations in the Context of HIV/AIDS in Malawi", Working Paper, University of Pennsylvania.
- Delavande A and H Kohler (2008b). "HIV Testing and Infection Expectations in Malawi", Working Paper, University of Pennsylvania.
- Dillon, B (2011). "Using Mobile Phones to Collect Panel Data in Developing Countries", *Journal of International Development*, forthcoming.
- Dominitz, J (1998). "Earnings Expectations, Revisions, and Realizations" *Review of Economics and Statistics* 80(3): 374-388.
- Dominitz, J and C Manski (1997). "Using Expectations Data to Study Subjective Income Expectations." *Journal of the American Statistical Association*

- ciation* 92(439): 855-867.
- Duflo, E (2001). "Schooling and Labor Market Consequences of School Construction in Indonesia: Evidence from an Unusual Policy Experiment." *American Economic Review* 91(4): 795-813.
- Ejrnæs, M and CC Portner (2004). "Birth Order and the Intrahousehold Allocation of Time and Education", *Review of Economics and Statistics* 86(4): 1008-1019.
- Engle, R (1982). "Autoregressive Conditional Heteroskedasticity with Estimates of the Variance of United Kingdom Inflation", *Econometrica* 50(4): 987-1007.
- Fafchamps, M (1993). "Sequential Labor Decisions Under Uncertainty: An Estimable Household Model of West-African Farmers", *Econometrica* 61(5): 1173-1197.
- Fafchamps, M and J Pender (1997). "Precautionary Saving, Credit Constraints, and Irreversible Investment: Theory and Evidence from Semi-Arid India", *Journal of Business and Economic Statistics* 15(2): 180-194.
- Filmer, D and L Pritchett (2001). "Estimating Wealth Effects Without Expenditure Data – or Tears: An Application to Educational Enrollments in States of India." *Demography* 38(1): 115-132.
- Giné, X, R Townsend and J Vickery (2008) "Rational Expectations? Evidence from Planting Decisions in Semi-Arid India", Working paper.
- Glewwe, P (2002). "Schools and Skills in Developing Countries: Education Policies and Socioeconomic Outcomes", *Journal of Economic Literature* 40(2): 436-482.
- Glewwe, P and M Kremer (2005). "Schools, Teachers and Educational Outcomes in Developing Countries", *Handbook on the Economics of Education*.
- Glewwe, P, M Kremer and S Moulin (2009). "Many Children Left Behind? Textbooks and Test Scores in Kenya" *American Economic Journal: Applied Economics* 1(1): 112-135.
- Hartley, M (1983). "Econometric Method for Agricultural Supply Under Uncertainty: Fertilizer Use and Crop Response", *Journal of Mathematical Analysis and Applications* 94: 575-601.
- Hildreth, C and JP Houck (1968). "Some Estimates for a Linear Model with Random Coefficients", *Journal of the American Statistical Association* 63: 584-595.
- Hill, RV (2010). "Liberalisation and Producer Price Risk: Examining Subjective Expectations in the Ugandan Coffee Market." *Journal of African*

- Economies* 19(4): 433-458.
- Jensen, R (2009). "The (Perceived) Returns to Education and the Demand for Schooling", *Quarterly Journal of Economics* 125(2): 515-548.
- Jolliffe, D (2004). "The impact of education in rural Ghana: examining household labor allocation and returns on and off the farm", *Journal of Development Economics* 73(2004): 287-314.
- Just, R and R Pope (1978). "Stochastic Specification of Production Functions and Economic Implications", *Journal of Econometrics* 7(1978): 67-86.
- Just, R and R Pope (2001). "The Agricultural Producer: Theory and Statistical Measurement", in B Gardner and G Rausser, eds., *Handbook of Agricultural Economics*, 631-735.
- Kiefer, N (2010). "Default Estimation and Expert Information", *Journal of Business & Economic Statistics* 28(2): 320-328.
- Lee, L-F and R Trost (1978). "Estimation of Some Limited Dependent Variable Models with Application to Housing Demand", *Journal of Econometrics* 8(1978): 357-382.
- Luseno, W, J McPeak, C Barrett, G Gebru and P Little (2003), "The Value of Climate Forecast Information for Pastoralists: Evidence from Southern Ethiopia and Northern Kenya", *World Development* 31(9): 1477-1494.
- Lybbert, T, C Barrett, J McPeak and W Luseno (2007). "Bayesian Herders: Updating of Rainfall Beliefs in Response to External Forecasts", *World Development* 35(3): 480-497.
- Manski, C (2004). "Measuring Expectations." *Econometrica* 72(5): 1329-1376.
- Marshall, A (1895). *Principles of Economics*. London: Macmillan and Co.
- McFadden, D, A Bemmaor, F Caro, J Dominitz, B-H Jun, A Lewbel, R Matzkin, F Molinari, N Schwarz, R Willis and J Winter (2005). "Statistical Analysis of Choice Experiments and Surveys", *Marketing Letters* 16:(3/4), 183-196.
- McKenzie, D, J Gibson and S Stillman. "A land of milk and honey with streets paved with gold: Do emigrants have over-optimistic expectations about incomes abroad?" World Bank working paper.
- Nerlove, M and D Bessler (2001). "Expectations, Information and Dynamics", in B Gardner and G Rausser, eds., *Handbook of Agricultural Economics*, 156-201.
- Nguyen, T (2008). "Information, Role Models and Perceived Returns to Education: Experimental Evidence from Madagascar", MIT working paper.

- Nyarko, Y and A Schotter (2002). "An Experimental Study of Belief Learning Using Elicited Beliefs", *Econometrica* 70(3): 971-1005.
- Papke, L and J Wooldridge (1996). "Econometric Methods for Fractional Response Variables With an Application to 401(k) Plan Participation Rates", *Journal of Applied Econometrics* II(1996): 619-632.
- Powell, J (1984). "Least Absolute Deviations Estimation for the Censored Regression Model", *Journal of Econometrics* 25(1984): 303-325.
- Press, W, S Teukolsky, W Vetterling and B Flannery (2007). *Numerical Recipes, 3rd Edition: The Art of Scientific Computing*. New York: Cambridge University Press.
- Psacharapoulos, G and H Patrinos (2002). "Returns to Investment in Education: A Further Update", *World Bank Policy Research Working Paper* 2881.
- Rosenzweig, M and K Wolpin (1993). "Credit Market Constraints, Consumption Smoothing, and the Accumulation of Durable Production Assets in Low-Income Countries: Investments in Bullocks in India", *Journal of Political Economy* 101(2): 223-244.
- Sahn, D and D Stifel (2003). "Exploring Alternative Measures of Welfare in the Absence of Expenditure Data", *Review of Income and Wealth* 49(4): 463-489.
- Santos, P and C Barrett (2010). "Persistent Poverty and Informal Credit", *Journal of Development Economics*, forthcoming.
- Schwarz, N, RM Groves and H Schuman (1998). "Survey Methods" in N Schwarz, RM Groves and H Schumann (Eds) *Handbook of Social Psychology, 4th edition*. New York: McGraw Hill, 143-179.
- Shen, H and Z Zhu (2008). "Efficient Mean Estimation in Log-Normal Linear Models", *Journal of Statistical Planning and Inference* 138: 552-567.
- Snyman, J (2005). *Practical Mathematical Optimization*. New York: Springer Science+ Business Media.
- Stoye, J (2010). "Partial Identification of Spread Parameters", *Quantitative Economics* 1: 323-357.
- von Neumann, J and O Morgenstern (1947). *Theory of Games and Economic Behavior*, 2nd. ed. Princeton: Princeton University Press.
- White, H (1980). "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity", *Econometrica* 48(4): 817-838.
- Wolpin, K (1987). "Estimating a Structural Search Model: The Transition from School to Work", *Econometrica* 55: 801-817.

Wooldridge, J (2002) *Econometric Analysis of Cross Section and Panel Data*.
Cambridge: MIT Press.