Title

Three Essays on Development Economics

Author

Ortiz Becerra, Karen

Publication Date

2022

Peer reviewed|Thesis/dissertation

Three Essays on Development Economics

By

KAREN ORTIZ BECERRA
DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Agricultural and Resource Economics

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

Ashish Shenoy, Chair

Michael Carter

Dalia Ghanem

Stephen Boucher

Committee in Charge

2022

To Juan Andrés and my 17-year-old self, dreams come true.

Acknowledgments

This dissertation would not have been possible without "my village" of mentors, teachers, family, allies, and friends. Thank you all for helping me achieve this goal.

I want to express my sincere gratitude to my advisors and committee members for all their support during these formative years. I am deeply grateful to Ashish Shenoy for encouraging me to study ambitious questions and providing valuable advice on many scholarly issues: from using structural models to address empirical questions to navigating data user agreements. I became a better and more confident researcher under your guidance. Many thanks to Michael Carter for his ever-insightful feedback and for helping me strengthen my ideas. Your work has had a meaningful impact on my research. Thanks also to Steve Boucher for keeping the doors of his office always open, for inspiring me to be a caring teacher, and for teaching me the value of fieldwork in research. I am grateful to Dalia Ghanem for being a caring mentor and role model and for all the practical and emotional support during this journey. Thank you for teaching me how to think formally about my ideas. It has been a privilege collaborating with you.

Many other scholars also had a significant influence on my career. I am grateful to Ana María Ibáñez and Raquel Bernal for believing in me and inspiring me to pursue this path. Thanks also to Travis Lybbert, Kristin Kiesel, and Marcela Eslava, whose examples taught me how to love teaching. I became an economist under the tutelage of Pierre Merél, Arman Rezaee, Marianne Bitler, Diana Moreira, Steve Vosti, Jim Wilen, Kevin Novan, Shu Shen, Andrés Carvajal, Takuya Ura, and many other excellent faculty. Many thanks to Mónica Parra, Daniel Mejia, and Sarojini Hirshleifer for showing me how to be a great collaborator.

My years in Davis have been a delight thanks to all the friends I made along the way: Oscar Barriga, Juan Correa, Cristina Chiarella, Aleksander Michuda, Matthieu Stigler, Armando Rangel, Francois Castonguay, Jessica Rudder, Laura Meinzen-Dick, Shri Kiruba, Andrea Estrella, Pierce Donovan, Tor Tolhurst, Joakim Weill, Stamatina Kotsakou, Caitlin Kieran, Miki Doan, Charlotte Ambrozek, Julian Arteaga, Marieke Fenton, Edward Whitney, Angelica Sauceda, Leonardo Caceres, and Lindsay and Michael Crawford. Cynthia van der Werf, thanks for your friendship and for providing perspective when I overthink things. I am also thankful to Jesus Arellano and Hovhannes Mnatsakanyan for the great company at the office and the many conversations about life, parenting, and research.

My parents and siblings have been essential to this accomplishment. Ma y Pa, todo lo que soy se los debo a ustedes. Mil gracias por su amor incondicional y todos los sacrificios que han hecho para apoyar mis sueños. Siempre sentí su amor, a pesar que estábamos lejos. Ana, you always knew how to lift my spirit when I needed it most. Our periodic check-ins became one of the highlights of my week. Juan, you are the

best “chief quality officer” any dissertation writer could ask for. Thanks for the help digitizing data and the steady stream of memes that always put a smile on my face. Johnathan, thanks for the frequent reminders about enjoying life and not taking everything so seriously. I always cherish the fun we have together.

I also want to express my appreciation for my extended family in Colombia. Thanks for your frequent notes of encouragement, for your prompt disposition to help during my fieldwork visits, for celebrating my accomplishments as your own, and for teaching me to appreciate life in the countryside. Our rural background, which I take pride in, is one of the main reasons I became interested in studying the development of rural economies. I hope that I can visit you more often. To my abuelita Elvia Murcia: thanks for transmitting the value of education through generations, even though your own dreams of studying were truncated by “La Violencia” at a very early age. You and my second abuelita, Elvia Romero, once dreamed of making a career in education. All the deliberate decisions you both made over the years have made it possible for me to fulfill those dreams.

There have been many other people who have helped me along the way. From the Carvajal Foundation, which funded a significant share of my tuition expenses at the Universidad de Los Andes, to incredible friends and kind administrative staff, who made it possible to juggle school responsibilities and enjoy my undergraduate studies while raising my son Juan. In a heroic act of generosity, Guillermo Camacho allowed me to bring Juan to the United States in order to pursue my dreams. I know how hard that decision was, and I will be forever grateful for that. I am grateful to my in-laws, Aiko and John Lee, for their overwhelming generosity, the invitations to explore the world together, and the stimulating conversations about history, inequality, and politics. Many thanks also to Carlos Castañeda for being a sounding board for my ideas and pushing me to think more deeply about public policy in Colombia. Pabis, thanks for your friendship, multiple visits, and putting up with me being busy all these years. I also want to thank you and the rest of our group of powerful girlfriends (Linis, Adri C., Adri M., Cami, and Mayis) for the inspiring chats, the recharging reunions, and the many memes that we have shared to support each other during the Ph.D.

My son and husband have been my rock during this journey.

To Juan: thanks for all your love and understanding of my busy schedule. You are my most important project and motivation. I am very proud of the young man you are becoming, the caring friend you are to many, and your passion and discipline with volleyball.

And finally, to Tim, my partner in all things: thanks for loving me and supporting me in every possible way. I am grateful for all the acts of service that lightened my workload at home, the encouraging words during setbacks, the TV shows you carefully curated to keep us laughing and trendy over these years, and your passion for learning new things about the world. You have made this accomplishment attainable and more rewarding. I am the lucky one.

Dissertation Abstract

This dissertation is comprised of two essays on development economics and one essay on methodological issues related to the impact evaluation of development interventions. The first two chapters explore the implications of farmland consolidation on rural development in Colombia. In the first essay, I estimate the effect of a previous consolidation event on the structural transformation of rural economies. In the second, I evaluate the impacts of potential consolidation policies on aggregate welfare. Finally, the last chapter analyzes how attrition affects the internal validity of field experiments and provides recommendations for current empirical practice.

Large-Farm Consolidation and Structural Transformation: Evidence from Colombia

The consolidation of farmland is accelerating in many developing countries. An important question that arises in this context is, what are the impacts of this consolidation on rural development? In the first chapter, I examine the effect of land consolidation on the structural transformation of rural Colombia. To motivate empirical work, I present a conceptual framework where the impact of consolidation on sectoral employment and wages depends on the strength of the pull response in the nonfarm sector relative to the push response in the farms. I examine this question by assembling a novel dataset of rural counties and leveraging quasi-experimental variation in response to a trade shock that changed the scale of production during the nineties. I find that counties with an increase in large-farm consolidation experienced a reallocation of labor from the agricultural to the nonagricultural sector. Yet, this labor reallocation led to a decline in workers' income over the medium term due to a sizeable increase in unemployment rates. These findings shed light on the implications of structural change within rural economies and the potential distributional impacts of consolidation across producers and workers.

Large-Farm Consolidation and Welfare in Rural Economies

Building on the first essay, the second chapter evaluates the impacts of large-farm consolidation on the aggregate welfare of rural populations. In particular, I develop a quantitative model of rural economies and conduct counterfactual experiments to evaluate the impacts of potential consolidation policies. This model features several empirical patterns connecting aggregate income with the distribution of farm sizes. The scale of operation affects agricultural productivity, while local wages and employment are determined by the concentration of profits through non-homothetic consumption growth. In line with previous work, I find that

large-farm consolidation increases the welfare of farmers due to gains in agricultural productivity. However, since the demand for rural labor decreases substantially, workers are adversely affected, and aggregate social welfare declines. I show that these effects vary by type of consolidation and are exacerbated when the rise of large operations is driven by merging the smallest farms. These findings shed light on the distributional impacts of consolidation and speak to the trade-off between productivity, employment, and social welfare inherent in land policies.

Testing Attrition Bias in Field Experiments

The third chapter shifts focus to methodological issues on the impact evaluation of development interventions. Randomized control trials are an increasingly important tool of applied economics. Non-response on outcome measures at endline, however, is an unavoidable threat to their internal validity. In this chapter, we approach the problem of testing attrition bias in field experiments with baseline outcomes. We differentiate between two internal validity questions. First, does the difference in mean outcomes between treatment and control respondents identify the average treatment effect for the respondent subpopulation? Second, is this estimand equal to the average treatment effect for the study population? For each object of interest, we establish identifying assumptions and propose procedures to test its sharp implications. We also document that the most widely used test in the field experiment literature, the differential attrition rate test, is not a valid test of internal validity in general and provide a Stata package to implement the procedures that we propose. These findings have public policy implications since some agencies use attrition rates to evaluate the reliability of research studies (e.g., What Works Clearinghouse, US Department of Education).

Contents

Chapter 1

Large-Farm Consolidation and Structural Transformation: Evidence from Colombia

1.1 Introduction

The reallocation of labor out of the agricultural sector as national income increases is a stylized fact of economic development. This process of structural change brought sustained productivity gains in early industrialized economies since technical progress enabled the exit of labor surplus in subsistence agriculture (Clark, 1940; Lewis, 1954; Kuznets, 1957). Currently, structural transformation is a central topic in the discussion of development policies in middle and low-income countries (Collier and Dercon, 2014; Barrett et al., 2017; Mcmillan et al., 2017; Mellor, 2017), where a large fraction of the population still lives in rural regions, and agricultural productivity is low.

One manifestation of structural transformation is the migration of workers from rural to urban markets. However, the reallocation of labor across sectors is also a key feature within rural economies themselves. Nonagricultural jobs account for 25% to 50% of rural employment in most developing countries (Lanjouw and Lanjouw, 2001). Additionally, nonfarm earnings are an increasingly important source of income diversification for those workers who are not fully absorbed in farm jobs (Haggblade et al., 2007, 2009; Foster, 2011).[1]

Despite the promising benefits of the structural change in rural economies, it remains unclear whether the

[1] In this paper, I use the terms nonagricultural and nonfarm interchangeably to facilitate exposition.

expansion of the nonfarm sector can boost local income and productivity. Nonfarm jobs are usually informal activities that yield a low return (Haggblade et al., 2007), and productivity gains relative to the farm sector are often modest (McCullough, 2017; Hamory et al., 2020). There is also wide spatial heterogeneity in the composition of this sector and its role in poverty reduction (Ravallion and Datt, 1999; Reardon et al., 2001). These empirical patterns complement the observation that structural transformation has different implications across space (Eckert and Peters, 2018) and suggest that the local nature and pace of labor reallocation determine how effective it is in promoting inclusive development.

A critical trend in developing economies over the past decades is the rise in the consolidation of farmland as a response to the boost in global agricultural demand. For one, large-scale transactions have taken place at an accelerated pace in low-income countries (Liao et al., 2020; Deininger et al., 2011). Meanwhile, middle-income countries already have more than 40% of their land on farms above five hundred hectares (Lowder et al., 2016). In the long term, this trend may lead to an increase in rural to urban migration, equalizing returns to labor across regions. In the short term, however, this consolidation can substantially affect the local returns to labor if there is a change in the demand for workers across sectors.

In this paper, I study the effects of large-farm consolidation on the structural transformation of rural economies. I focus on two main questions of interest. First, does the consolidation of land affect the local reallocation of labor across sectors? Second, do rural workers benefit from this consolidation?

To motivate this empirical work, I lay out a conceptual framework that links land consolidation with the aggregate demand for workers in rural economies.[2] A key feature of this framework is that consolidation simultaneously affects the demand for labor in both economic sectors. On the one hand, labor intensity impacts the demand for farm labor. On the other, nonfarm labor demand is affected by non-homothetic consumption growth. The main insight of this framework is that consolidation leads to a push response in the demand for labor out of the farm sector and a likely pull response in nonfarm labor demand. Thus, if the pull response is small relative to the push response in the farm sector, consolidation may lead to the reallocation of workers towards the nonfarm economy along with a reduction in their wages.

The Colombian setting offers an excellent opportunity to examine these questions since there has been a significant change in overall land concentration over time. For instance, between 2000 and 2012, the Gini index of landholdings increased from 0.854 to 0.872 (IGAC, 2012). Furthermore, given the diversity in topography and agricultural suitability, this change in concentration has been widely heterogeneous across space. This setting is also relevant from a policy perspective, given the upcoming reform aiming to redistribute land to landless peasants and farmers with insufficient acreage (Arteaga et al., 2017). Besides, a large majority

[2]This framework draws on insights from the quantitative model developed in Ortiz-Becerra (2022) to analyze the impacts of consolidation on the welfare of rural populations.

of rural households derive their income from labor markets due to their lack of access to land (58%).

I empirically examine the impact of a specific shift in land concentration that occurred during the 1990s. This shift was largely driven by a change in the terms of trade that transformed land use in the country. The area in pastures and land-intensive crops increased by more than two million hectares, while coffee and other labor-intensive crops lost important participation in agricultural land (Balcázar, 2003). Overall, the gini of landholdings displayed a slight increase between 1993 and 2005 and large-farm consolidation changed in almost all the municipalities.

I study this shift in land concentration using a novel panel dataset of rural municipalities with measures on employment and land consolidation. This dataset links information from different sources, including the population census, the national household surveys, and the rural cadastre. I examine impacts on the reallocation of labor across sectors and workers' income using measures of wages and unemployment rates, and define large-farm consolidation as the share of area in large farms in each municipality.

To estimate treatment effects, I leverage quasi-experimental variation in the ability of local economies to respond to land-use changes driven by trade liberalization. Specifically, I construct an instrument for the change in large-farm consolidation based on topographic features of the terrain that influence the financial viability to produce at large scale. Consolidation is negatively correlated with the terrain's degree of inclination since it is cheaper to produce at large-scale in flatter grounds due to infrastructure's construction costs and cattle carrying capacity.

I account for two main factors that might be correlated with a town's topography and represent a threat to the exclusion restriction of the instrument. First, the differential effects in agricultural production and labor force participation driven by the upsurge of conflict in rural areas (Fernández et al., 2014; Arias et al., 2018). Second, labor market trends associated with crop suitability, including the potential income effects after the shift in the terms of trade. For instance, changes in tariffs and relative prices were more likely to benefit economies that had an *absolute* advantage to produce exportable crops such as flowers and oil palm. My analysis compares adjacent towns with similar agricultural suitability but distinct feasibility to produce at scale, so the variation I exploit arises from the ability to consolidate rather than the *absolute* advantage to produce a particular crop.

I find that the rural economies with an increase in large-farm consolidation experienced a reallocation of labor out of the agriculture sector. Specifically, a one standard deviation increase in consolidation resulted in a ten percentage point decline in the share of farmworkers. In addition, consolidation led to a fourteen percentage point increase in the unemployment rate with no significant change in wages, suggesting that this reallocation of labor was accompanied by a net negative impact on workers' income. My estimates imply that one standard deviation increase in land consolidation - corresponding to a twelve percentage point rise

in the share of area in large farms– explains 70% of the observed decline in farm employment for the *average* Colombian municipality between 1993 and 2005.

I provide evidence that these findings are not driven by several factors that may confound the relationship between employment and the instrument. First, parallel pre-trends support the exclusion restriction that topography does not lead to differential trends in labor markets across similar towns. Second, results are unchanged when controlling for the upsurge of conflict and forced displacement during this period. These findings are also robust to the use of different inference procedures, including alternative types of intra-cluster correlation and size corrections for weak instruments.

Taken together, my findings indicate that consolidation led to a decrease in the demand for farmworkers and an absolute reduction in the demand for rural labor. These results are consistent with the predictions of the conceptual framework when the pull response in the nonfarm sector is small relative to the push response in the farms. This decline in labor demand, however, appears to be smaller in economies where expenditure in local markets is expected to be large, suggesting that non-homothetic consumption growth is an important mechanism behind this effect.

My results on unemployment suggest that labor mobility across space was not large enough to offset the decrease in labor demand after consolidation.[3] These findings are consistent with the low levels of integration across rural labor markets during the 1990s (Nupia, 1997) and the prevalence of different barriers that limit migration in developing contexts (Morten and Oliveira, 2018; Dix-Carneiro and Kovak, 2017; Bryan et al., 2014). Since my period of analysis spans twelve years, an important implication of these results is that land consolidation has lasting effects on local labor markets. This is in line with previous empirical work that examine the capacity for labor markets to respond to economy-wide shocks and document persistent effects in the short and medium run (Dix-Carneiro and Kovak, 2019; Dix-Carneiro, 2014; Autor et al., 2014; Artuç et al., 2010). For instance, Dix-Carneiro and Kovak (2019) find that the effect of Brazil's trade liberalization on local unemployment and informality persisted over ten years.

This paper contributes to the literature on structural transformation (Herrendorf et al., 2014). My findings show that land consolidation is another driver of the reallocation of labor across sectors. Recent research studies the impacts of trade on structural transformation (Farrokhi and Pellegrina, 2020; Fajgelbaum and Redding, 2018; McArthur and McCord, 2017), and suggests that land inequality and input use are important mechanisms through which price booms affect labor reallocation (Laskievic, 2021). Yet, while these papers exploit the differential exposure to shocks in order to estimate the direct effects of trade on

[3]This is not to say that the extent of rural-out migration was small during this period. On the contrary, due to the upsurge of conflict, a large proportion of rural households were forced out of rural communities. In my analysis, however, forced displacement is a potential threat to internal validity. Thus, my empirical strategy compares municipalities with a similar incidence of this type of labor mobility. Future access to net migration flows will allow me to assess how much of the observed changes are driven by economic migration.

structural change, I examine the direct impact of large-farm consolidation by leveraging the isolated effect of market integration on the scale of agricultural production across towns with different topography and similar trends in trade exposure.

This paper also complements recent studies that examine the spatial implications of structural transformation (Bustos et al., 2020; Eckert and Peters, 2018; Nagy, 2016; Desmet and Rossi-Hansberg, 2014; Michaels et al., 2012; Caselli and Coleman II, 2001) by showing that, in rural settings, this transformation can be accompanied by a decline in workers' income over the medium term. This result implies the opposite relationship between income and labor reallocation at a national level, and is consistent with previous work on the delayed response of local labor markets to economy-wide shocks (Dix-Carneiro, 2014; Dix-Carneiro and Kovak, 2019). Methodologically, my empirical approach is also similar to previous work that exploits quasi-experimental variation to examine the determinants of structural change at a local level rather than across regions (Bustos et al., 2020; Uribe Castro, 2020; Bustos et al., 2016; Foster and Rosenzweig, 2008).

Finally, my work contributes to two main strands of the long-standing literature on land allocation and economic development by focusing on local labor markets effects. First, I add to the growing set of empirical studies that examine the impacts of large-scale transactions on rural welfare (Liao et al., 2020; Ali et al., 2019; Deininger and Xia, 2016). With a few exceptions, these studies rely on cross-sectional data and unconfoundedness to estimate spillover effects on smallholders (Bottazzi et al., 2018; Herrmann, 2017; Jiao et al., 2015). In contrast, I provide novel estimates of the economy-wide impacts on aggregate employment and wages and relax the assumption of unconfoundedness by leveraging a 2sls approach along with panel data. There is also a wide body of theoretical work that examines the relationship between scale and agricultural productivity and income (Eswaran and Kotwal, 1986; Adamopoulos and Restuccia, 2014; Foster and Rosenzweig, 2017; Adamopoulos et al., 2019). This paper adds to that work by showing that land consolidation can decrease the income of workers despite these potential productivity gains.

The rest of this paper is organized as follows. The next section lays out the conceptual framework and its main analytical insights. In Section 2.2, I provide details on the setting and the data. Section 1.4 describes the empirical strategy to estimate the impacts of large-farm consolidation. Section 1.5 discusses the results and potential mechanisms. Section 1.6 concludes.

1.2 Conceptual Framework

This section discusses two main opposing mechanisms through which large-farm consolidation affects the structural transformation of rural economies. This conceptual framework illustrates how land consolidation simultaneously affects the demand for workers in the agricultural and nonagricultural sectors, drawing on

insights from the two-sector model developed in Ortiz-Becerra (2022).[4]

An empirical regularity in agrarian economies is that there is an inverse relationship between labor intensity and landholding size. This observation has long been documented in several developing contexts (Sen, 1981; Carter, 1984), and more recently it has been shown to be accentuated over time (Deininger et al., 2016). Some of the reasons behind this relationship are factor market imperfections that constrain smallholders in their ability to increase their scale of operation, substitute own for hired labor, and/or substitute labor for other inputs. For instance, there is ample evidence that agricultural machinery saves on labor costs (Hornbeck and Naidu, 2014; Davis, 2017) and that mechanization is more likely to occur on larger farms (Foster and Rosenzweig, 2017). This pattern suggests that large-farm consolidation can reduce farm labor intensity and lead to a decline in the demand for workers in the agricultural sector.

Land consolidation can also affect the demand for nonfarm labor through changes in the demand for locally produced goods. One one hand, scale production can increase aggregate agricultural productivity and income (Adamopoulos and Restuccia, 2014; Foster and Rosenzweig, 2017). On the other, the expenditure share in local consumption declines with income since wealthier individuals are less likely to spend their money in local markets. Consider, for example, the case in which individuals prefer to buy clothing and jewelry from the city as their income increases. These observations suggest that the consumption preferences are non-homothetic, and hence, the demand for rural nonfarm workers depends on how concentrated agricultural profits are.

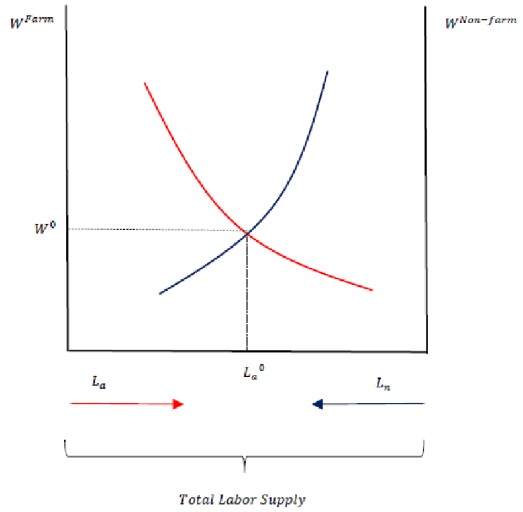
To illustrate how these two primary forces affect wages and labor allocation across sectors, consider the rural labor market depicted in Panel A of Figure 2.4. An increase in land consolidation leads to a push response in the demand for farm workers (red line) as agriculture becomes less labor intensive. Meanwhile, the shift in the demand for nonfarm workers (blue line) is ambiguous and depends on how much of the gains in agricultural productivity are spent on local consumption. Overall, the net effect of large-farm consolidation on workers' income depends on whether the pull response in the nonfarm sector is larger than the push response in the farms. If the gains in productivity lead to a substantial increase in the demand for nonfarm labor, the equilibrium wage rises (Panel B). In contrast, if there is a reduction in the demand for workers in this sector, land consolidation leads to a wage decline (Panel C).

In this paper, I estimate the overall impact of large-farm consolidation in the Colombian setting using quasi-experimental variation in the scale of production across a broad set of rural municipalities. In addition, I provide empirical evidence suggesting that these effects are largely driven by changes in local labor demand.

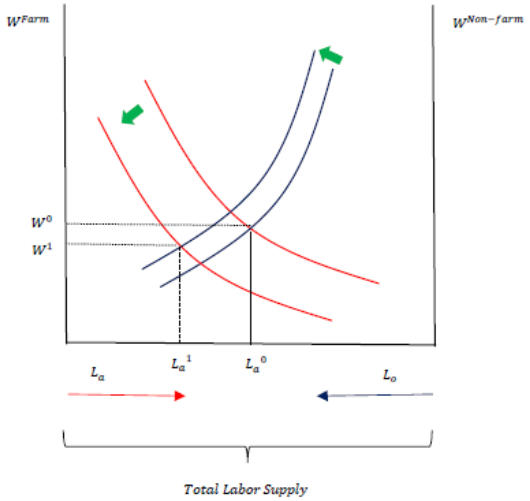
[4]This static model represents a benchmark economy with a fixed supply of labor and constitutes a useful point of reference to analyze impacts in the short term.

Figure 1.1: Large-Farm Consolidation and Equilibrium Wage: An Illustration

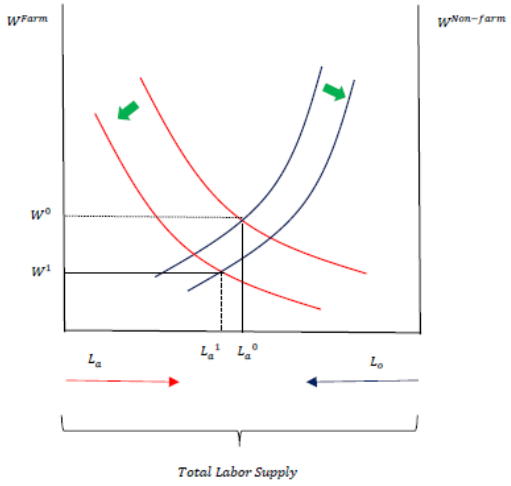
Panel A. Rural Labor Market



Panel B: Positive Pull Effect



Panel C: Negative Pull Effect



Notes: Own illustration.

1.3 Context and Data

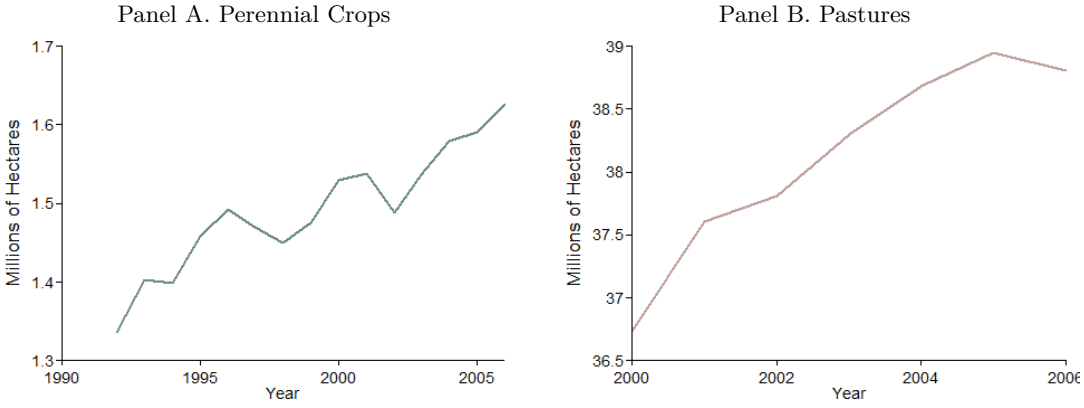
1.3.1 Trade Liberalization and Productive Transformation in Agriculture

Colombia underwent a gradual process of trade liberalization in the early 1990s. This liberalization ended an import substitution regime that lasted over forty years. During this decade, the average tariff in the whole economy declined from 38.6% to 11% (Jaramillo, 1998). At the same time, the country completed the negotiation of its first free trade agreements and implemented policies to promote exports.

Agriculture was one of the primary industries targeted by this trade liberalization. In particular, the average tariff in the sector declined from 31.5% to 15%. In addition, the government implemented several policies to promote integration with the international market. Some policies included subsidies and tax exemptions for the production of perennial cash crops with apparent competitive advantages (e.g., flowers, bananas, oil palm, sugar cane). Other policies aimed to mitigate the impacts of international competition on the production of seasonal crops such as soy, wheat, and barley.[5]

These policies led to a change in the profitability of crops and a productive transformation of the agricultural sector (Balcázar, 2003). On the one hand, the production of perennial crops increased as a response to the new market opportunities and the rise in profitability. On the other, the production of tradable seasonals declined due to the higher competition with international producers. Figure 1.2 shows the evolution of land use in the country during this period. Despite the decrease in aggregate cultivated area, the area in pastures and cash crops increased by more than two million hectares between 1993 and 2005.

Figure 1.2: Area in Land-Intensive Crops & Pastures 1993-2005



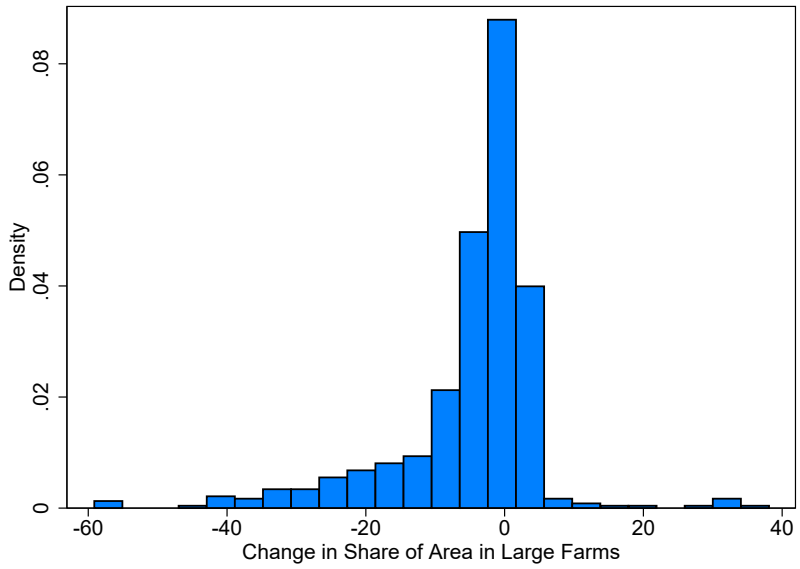
Notes: Own illustrations using data from the Municipal Agricultural Evaluations.

Thus, there was a change in the scale of agrarian operation and land consolidation in the country. In

[5]Some of the policy measures that were implemented to promote agricultural exports were the Law 693 of 2001, the Law 818 of 2003, and the creation of the Price Stabilization Fund.

some regions, farms were consolidated to produce land-intensive cash crops such as cattle, flowers, and oil palm.[6] In others, farms were fragmented to mitigate the impacts of the deterioration in the terms of trade. These two countervailing forces led to a wide geographical variation in the number and size of large farms. As a result, nearly 20% of the towns displayed an increase in large-farm consolidation while 30% experienced a decline below the national average of five percentage points (see Figure 1.3).[7]

Figure 1.3: Change in Large-Farm Consolidation, 1993 -2005



Notes: This figure displays the change in the share of area in large farms (in percentage points) for rural municipalities between 1993 and 2005. Own calculations using cadastral data.

1.3.2 Data Sources and Measurement

I gather data from multiple sources and put together a dataset of municipalities for the years 1993 and 2005. Municipalities are the smallest administrative units in the country and are a good approximation to local labor markets since most of their inhabitants work within their geographic boundaries.[8] I focus on rural economies and exclude large cities and their main agglomerations using the definition proposed by the *Rural Mission* in 2014.[9] To account for the segregation and creation of new municipalities across time, I use a consistent unit of observation based on the official boundaries of the municipalities in the year 1993. In my

[6]These crops are usually produce at large-scale due to a high degree of vertical integration. Furthermore, cattle has a maximum carrying capacity per hectare that requires a minimum plot size threshold to render production profitable.

[7]Note, however, that despite this average decline in large-farm consolidation, overall land concentration (as measured by the Gini index) increased. This observation suggests that there was an increase in the dispersion of farm sizes across the whole distribution after the market integration, which is consistent with recent evidence on the impact of price booms on the Gini of land in similar contexts (Laskievic, 2021).

[8]According to the population census of 2005, 95% the workers who lived in rural municipalities worked in the same municipality where they lived.

[9]This definition classifies municipalities in four categories based on population density and the number of inhabitants living in the town's seat. See Ocampo (2014) for more details.

final sample, only 4% of the units of observation are composed of two or more adjacent municipalities.

Unemployment and sectoral employment: I use information from the National Population Census of 1993 and 2005 to construct these measures. These two rounds of the census collect information on labor participation and the industry of employment, in addition to the usual demographic and socio-economic characteristics collected in other rounds.[10] For each municipality, I compute the unemployment rate as the ratio of the number of unemployed individuals to the total population that is economically active. Similarly, I construct the share of agricultural workers using information on the economic sector of the main job in the last 15 days. I define agricultural jobs as those activities that make an intensive use of land to produce crops and raw materials, raise livestock and poultry, or farm fish. Any other activities, including the processing of food and beverages, are classified as non-agricultural jobs.[11]

Wages: Since the population census does not collect data on wages, I use individual-level data from the National Household Surveys (ENH) of 1998 and 2009. These repeated cross-sections are representative at the national and department level and include a large set of rural municipalities that are selected at random each round. The employment module of this survey collects data for a random sample of workers on features such as occupation, hours worked, and income. I compute the hourly wage for each worker as the ratio of her monthly wage to the number of hours worked during the last month, and convert nominal to real values using the price index of 1993. For all outcomes, I focus on individuals between 15 and 65 years old.

Land consolidation: I use data from the national cadastre system for the years 1993 and 2005 to construct measures of land consolidation.[12] This system is a census of all the properties in the country with detailed information on the location of the plot, the type of holder, and the plot's size.[13] For each municipality and year, I have information on the number of properties and their total area across thirteen size ranges.[14] To focus on privately held land, I exclude the records of indigenous reserves and properties that belong to the State.

Two main features make this data advantageous to construct measures of concentration and consolidation. First, the system records information based on possession instead of ownership. Therefore, the records of

[10]These two rounds of the Population Census have been shown to be comparable despite their differences in implementation (Le Roux, 2013; Mallarino, 2007; Jaramillo and Ibáñez, 2005). Access to this data was obtained per confidential agreement with the National Department of Statistics.

[11]Two main reasons indicate that the timing of data collection does not primarily drive the differences in agricultural employment across locations. First, in most regions, the share of agricultural workers does not vary much throughout the year (see Figure 2A in the appendix). Second, according to experts at the National Department of Statistics, the roll-out of the censuses was not correlated with any particular season.

[12]The data from 2005 was purchased from the National Institute of Geographic Information (IGAC), and the data from 1993 was generously provided by Fabio Sánchez at the University of the Andes in Colombia.

[13]The national cadastre is managed by five different agencies: the National Institute of Geographic Information (IGAC), the department of Antioquia, and the capital cities of Bogota, Medellin, and Cali (IGAC, 2012). In this analysis, I use the information from Antioquia and IGAC, which constitutes the whole universe of rural municipalities with cadastral data.

[14]These size ranges are: less than 1 hectare, 1 to 3 hectares, 3 to 5 hectares, 5 to 10 hectares, 10 to 15 hectares, 15 to 20 hectares, 20 to 50 hectares, 50 to 100 hectares, 100 to 200 hectares, 200 to 500 hectares, 500 to 1000 hectares, 1000 to 2000 hectares, and more than 2000 hectares.

private land include farmers with informal titles or settled in vacant public lots. Second, the measures of plot size are less likely to be afflicted by self-reporting bias – common in survey data– as the data gathered upon the *cadastre formation* is collected and updated by personnel in the field (IGAC, 1988).[15]

I define large-farm consolidation as the share of area in large farms. To determine what a large plot is for each town, I use as reference the average family farm unit, a policy instrument representing the minimum plot size needed to generate an income surplus given the agro-ecological conditions of the plot's location.[16] Following Machado and Suarez (1999), I use a threshold of ten units to characterize large plots. Therefore, large-farm consolidation for each municipality i is given by: $\omega_i^{large} = \omega_i^a \quad if \quad \kappa_i \in [a, b)$, where κ_i refers to ten times the family farm unit, (a, b) refer to the lower and upper bounds of the observed size ranges, and ω_i^a refers to the share of area in plots with a size of at least a hectares.[17] In addition to this measure, I also use this data to calculate the Gini index of landholdings, which is a measure of *overall* concentration across the whole plot distribution.

Additional municipal features: I compile a set of municipality characteristics using different data sources. First, I obtain information on administrative divisions and the town's average family farm unit from the National Department of Statistics (DANE). Second, I use administrative data from the National Memory Center to calculate measures of conflict intensity and the prevalence of forced displacement between 1993 and 2005.[18] Third, I digitize data on initial levels of urbanization, such as the total number of inhabitants and the share of the rural population, using the 1985 census. Finally, I calculate topographic measures on elevation and the average degree of terrain's inclination using the Data Elevation Model from NASA and the shapefiles of municipalities with 1993 boundaries. In contrast to other sources, these topographic measures are very precise as the pixel resolution is one arc-second (approximately 30 meters). This is particularly relevant for my analysis since it implies that the instrument I construct for land consolidation displays a considerable variation across towns.

[15] One important caveat of this data, however, is that resource limitations prevent cadastral agencies from carrying out the updates during the established 5-year window for all the municipalities. In fact, Pinzón and Fonti (2007) find that only 31% of the municipalities were fully up-to-date in 2005. Yet, to the extent that the updates are not correlated with municipality's characteristics in general (Martinez, 2020), it is unlikely that my estimates are largely driven by differences in measurement error.

[16] The family farm unit was initially created by the Law 135 of 1961 to guide the allocation of vacant public lands. It represents the minimum plot size required to produce an income of three monthly minimum wages and a disposable income after paying land rent payments. This unit takes on different values depending on the type of agro-ecological zone and land use. Thus, the average unit for each town is calculated as the weighted mean across production systems and zones (Departamendo Nacional de Planeación, 2000).

[17] For instance, if the average family farm unit for the town is 32 hectares, my measure of large-farm consolidation equals the share of area in plots of at least 20 hectares. Since this measure based on size ranges is less precise than the one I would obtain with microdata, I plan to check whether the main results of the paper are robust to different definitions of ω_i^{large} in a future version of this draft.

[18] These indicators include the share of the forcefully displaced population, the number of murdered individuals, and the total number of attacks, incursions, and assaults on the civilian population perpetrated by illegal armed groups. For more details on this data, see Centro Nacional de Memoria Historica (2013).

1.3.3 Sample of Rural Municipalities

The study population in this analysis consists of 590 municipalities across 21 *departments* of the country. These 21 departments correspond to 50% of the country's area and 95% of the national population.[19] My study sample accounts for three-fourths of the rural economies in these departments.[20] These towns are spread in similar proportions across lowlands, hills, and highlands, thereby representing the main agro-climatic regions in the country.

In Table 1.1, I present summary statistics that describe the main characteristics of the sample before the observed changes in land concentration (i.e. 1993). These municipalities are located about 79 kilometers from the department's capital and have a mean area of 537 square kilometers. In contrast to large cities, the population count was low and the the majority of workers worked in agricultural activities. These rural towns also had a low unemployment rate and a wide spread in the wage perceived by workers. While the average hourly wage was slightly above the minimum legal wage at COL$524, close to two-thirds of the workers were making less than that amount.[21]

Table 1.1: Main Characteristics of The Study Population in 1993

	N	Mean	S.D.	Min	p25	p75	Max
Time-invariant characteristics							
Total area (km2)	590	536.7	1,121.7	20.0	127.0	541.0	17,536.0
Distance to departmen's capital (km)	590	78.7	47.7	9.3	44.2	102.8	276.0
Family farm unit (ha)	590	22.0	14.1	2.5	13.1	27.3	125.6
Terrain's inclination (degrees)	590	29.7	14.5	1.5	19.2	40.7	64.8
Time-variant characteristics (1993)							
Total population	590	15,543	14,865	1,277	6,389	19,865	171,936
Share rural population (%)	590	68.7	18.5	4.8	58.4	82.8	97.2
Share of farm employment (%)	590	66.6	17.5	0.7	56.6	80.7	95.9
Unemployment rate (%)	590	2.5	2.2	0.0	1.0	3.5	17.9
Hourly wage (1993 COL$)†	1,366	524.1	613.5	3.4	102.2	687.5	6642.9
Gini of landholdings	590	0.38	0.16	0.04	0.26	0.51	0.90
Share of area in large farms (%)	590	29.1	21.4	0.0	11.3	42.4	99.4

Notes: (†) The statistics on hourly wage refer to the wage workers in the 1998 National Household Survey. The average hourly wage of $524 pesos in 1993 corresponds to $3,806 pesos in 2019, which is approximately equivalent to an hourly wage of U$1. See Sections 2.2.1 for details on data sources and definitions.

Finally, the size threshold that defines a large farm has a mean of 220 hectares and varies widely across towns. For instance, while in some municipalities, a large farm has at least 800 hectares, in others, farms of 25 hectares are considered large. The share of area in large farms across municipalities was 29% on average and one-quarter of these towns had a share above 42%.

Table 1.2 reports the changes in consolidation and rural employment in my sample.

[19] I exclude the municipalities from the departments of Amazonas, Arauca, Casanare, Caquetá, Chocó, Guainía, Guaviare, Putumayo, San Andrés, Vaupés, and Vichada for at least one of the following reasons: i) department was not part of the agricultural frontier in the 1990s, ii) rural areas consisted mostly of indigenous reservations and afro-colombian lands, iii) cadastral data was either non-existent or unreliable.

[20] The remaining 25% of rural towns do not have complete data for both years of analysis.

[21] An hourly wage of COL$524 in 1993 is equivalent to COL$3,806 per hour in 2019 (U$1 per hour).

Table 1.2: Changes in Main Variables 2005- 1993

	1993		$\Delta_{2005-1993}$	
	Mean	SD	Mean	SD
Share of agricultural workers	0.666	0.174	-0.092	0.138
Unemployment rate	0.025	0.022	0.080	0.123
Gini of landholdings	0.386	0.157	0.006	0.079
Share of area in large farms	0.291	0.214	-0.059	0.129
Number of conflict-related events since 1993	1.017	1.951	22.692	31.833
Number of homicides since 1993	148.33	265.73	346.63	509.04
Number of displaced individuals since 1993 (expelled)	311.71	820.02	3505.02	7802.53
Number of displaced individuals since 1993 (received)	93.58	307.40	1850.23	4318.27

Notes: The number of municipalities included is 590. See Section 2.2.1 for details on data sources and definitions.

1.4 Empirical Strategy

To examine the impacts of large-farm consolidation, I use a 2sls approach that exploits the differential ability of the municipalities to respond to this trade-induced shock. In particular, I construct an instrument for the change in consolidation based on topographic features of the economy that influence the financial viability of production at a large scale. This instrument is defined as the average degree of inclination for the terrain in each municipality.[22] In my sample, the degree of inclination ranges from 1.5° to 64.8° and has an average of 29.7°.

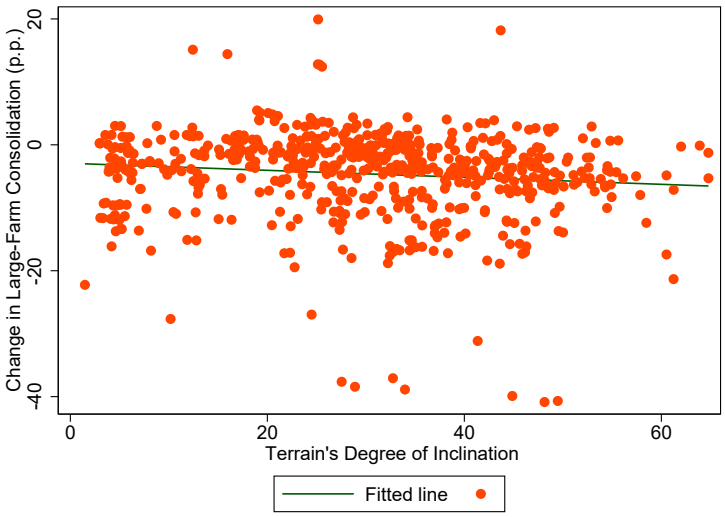
The rationale behind this instrument is that the construction and maintenance of infrastructure for large-scale production are cheaper in flatter terrains. For instance, export crops such as flowers and oil palm require greenhouses and irrigation systems. Likewise, flat landscapes are more attractive for mechanization and extensive cattle ranching, as the carrying capacity per hectare decreases with the gradient of inclination.[23] Declines in large-farm consolidation are also smaller in locations with lower degree of inclination, as the opportunity cost of fragmentation is higher. For instance, while coffee farms in the slopes are fragmented after declines in the international coffee price, coffee farms in the inter-Andean valleys are often transformed into cattle farms (Balcázar, 2003; García, 2003). This negative relationship between large-farm consolidation and the economy's degree of inclination is depicted in Figure 1.4.

A town's topography also influences crop suitability. Thus, one important threat to the exclusion restriction of this instrument is that locations with different crop portfolios faced distinct income and labor market shocks due to trade liberalization. For instance, economies that were suitable to produce cash crops experienced an improvement in their terms of trade, while economies that produced importable crops were negatively affected in terms of profitability. To address this concern, I compare contiguous municipalities

[22]This measure corresponds to the area-weighted average across all the raster cells that intersect with the municipality's surface. For more details on the data used to construct this measure, see Section 2.2.1.

[23]According to interviews in the field, the carrying capacity of cattle in flatlands is about three cows per hectare compared to two cows (or less) in the slopes.

Figure 1.4: Change in Large-Farm Consolidation & Terrain's Inclination



Notes: This figure depicts the fitted values of a regression of the change in consolidation on the terrain's inclination and province fixed effects.

of the same province (δ_p), which share similar crop suitability and are subject to the same governmental policies.[24]

Similarly, previous evidence suggests that the upsurge in conflict during this period likely had differential effects on labor markets of economies with distinct topography. On the one hand, conflict induced changes in agricultural production and labor supply (Arias et al., 2018; Fernández et al., 2014). On the other, the upsurge was more pronounced in the mountains as they provide a natural shelter for illegal groups (Centro Nacional de Memoria Historica, 2013). I account for this potential relationship by controlling for changes in forced displacement and trends in conflict-related measures such as homicides and the number of attacks (ΔC_{mp}).

For the analysis on unemployment and sectoral employment, I estimate equations of the form:

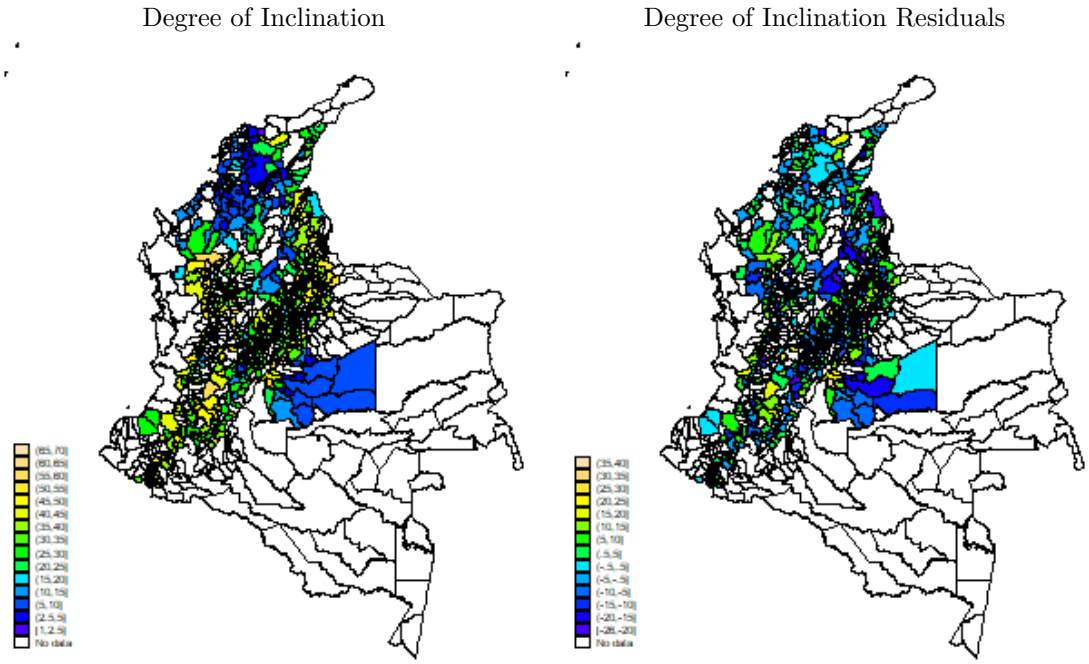
$$\begin{aligned} \Delta Y_{mp} &= \beta \Delta \hat{D}_{mp} + \delta_p + \alpha_1 \Delta C_{mp} + \Delta u_{mp} \\ \Delta D_{mp} &= \gamma S_{mp} + \delta_p + \alpha_2 \Delta C_{mp} + \Delta \eta_{mp} \end{aligned} \tag{1.1}$$

where ΔY_{mp} refers to the change in the outcome of interest between 1993 and 2005 for municipality m in province p, ΔD_{mp} refers to the change in large-farm consolidation, and S_{mp} refers to the terrain's degree of inclination. In all estimations, standard errors are clustered at the municipality level to account for serial correlation. To give an idea of the variation I use to estimate β, I show the spatial distribution of the

[24]Provinces are department subdivisions that have been historically used to plan and develop environmental and territorial policies (DANE, 2014). There are 100 provinces in my sample with an average of eight municipalities per province.

instrument and its values after partialling out the fixed effects and covariates in Figure 1.5.

Figure 1.5: Spatial Variation in Terrain's Inclination



Notes: These maps display the spatial distribution of the municipality's terrain inclination (in degrees) and its values after partialling out province fixed effects and the change in conflict-related controls 1993-2005.

For the analysis on wages, I use repeated cross-sections of workers in rural municipalities. Thus, I estimate the following modified version of Equation 1.1:

$$\begin{aligned} Y_{imjt} &= a_m + \beta \hat{D}_{mjt} + \delta_{j,t} + \alpha_1 C_{mjt} + \lambda_1 X^{85}_{m,t} + u_{imjt} \\ D_{mjt} &= a_m + \gamma(S_{mj} \times T_t) + \delta_{j,t} + \alpha_2 C_{mjt} + \lambda_2 X^{85}_{m,t} + \eta_{mjt} \end{aligned} \tag{1.2}$$

where Y_{imjt} refers to the log hourly wage of worker i in municipality m, at time t, and T_t is a dummy variable that takes the value of one (zero) for the year 2009 (1998). These cross-sections include the subset of rural municipalities that were sampled in both years of the survey. Since only a few of these towns belong to the same province, I use department-year fixed effects to account for income changes induced by the change in the terms of trade ($\delta_{j,t}$). Hence, I also control for initial urbanization rates ($X^{85}_{m,t}$) to account for differential trends in development across towns of the same department.

In contrast to the analysis in Equations 1.1, I use a two-sample 2sls approach to estimate the Equations in 1.2.[25] This approach allows me to estimate the *first-stage* with the full sample of municipalities to improve

[25]The two-sample 2sls approach was proposed by Angrist and Krueger (1995) as a procedure to obtain 2sls estimates using two independent samples; one for each stage. For more information on the consistency of this procedure, see Inoue and Solon (2010), Pacini and Windmeijer (2016), Choi et al. (2018), and Angrist and Krueger (1992).

the precision of these estimates.[26] I follow Pacini and Windmeijer (2016) to correct the standard errors of this two-step procedure and obtain estimates that are robust to heteroskedasticity.

The exclusion restriction in Equation 1.1 is that municipality's terrain inclination does not lead to different labor market trends across contiguous economies within the same province.[27] This assumption would be violated if the design and execution of public policies in a province depends on topography, or if the trade-induced income effects - associated with the production of certain crops- are not fully accounted for. In Table 1A, I show that rural municipalities with distinct levels of terrain inclination had similar trends in urbanization and population growth before trade liberalization.[28] While the exclusion restriction in untestable, these results provide reassuring support for this assumption.

In this case, β identifies the average marginal effect on the subpopulation of economies that respond to encouragement from the instrument (i.e. compliers).[29] This subpopulation comprises the towns where the shift in land consolidation was driven by the physical and financial ability of production at scale.

In the case of Equation 1.2, the exclusion restriction is stronger as the larger variation in agro-climatic conditions across municipalities of the same department imply that crop-specific income effects are not fully accounted for . In line with this, Table 1B shows that towns with different terrain's inclination within the same department displayed differences in population growth before trade liberalization. Therefore, the results of the analysis on wages should be regarded with caution.

1.5 Results

Table 1.3 shows the results on the structure of rural employment and unemployment rates. Panel A reports the first stage relationship and the reduced form effects of the municipality's inclination gradient. Panel B reports the corresponding 2sls estimates for the outcomes of interest and the Kleibergen-Paap F-statistic that tests for the first-stage.[30] I also report weak instrument-robust confidence sets to avoid bias due to pre-testing for the instrument's relevance (Andrews et al., 2019).

The first-stage results confirm the negative relationship between the terrain's inclination and the change

[26] Naturally, the second stage is estimated using only the subset of towns with data on workers' wage.

[27] Formally, $E(s_{mp}\Delta u_{mp}|\Delta C_{mp}, \delta_p) = 0$. This assumption is weaker than the one required for the within estimator, which states that the *idiosyncratic* component of land consolidation across municipalities of the same province is not correlated with labor market shocks; i.e. $E(\Delta D_{mp}\Delta u_{mp}|\Delta C_{mp}, \delta_p) = 0$.

[28] These two variables are closely related to changes in sectoral employment and labor supply in a local economy. I will analyze pre-trends on my outcomes of interest once I am granted re-access to the microdata of the 1985 population census at the Colombian National Department of Statistics.

[29] Monotonicity is an additional assumption that is required to interpret β as the local average marginal effect of land consolidation on the outcomes of interest. This assumption, untestable in nature, implies that while increasing the average slope can either encourage land consolidation or have no effect at all, it cannot discourage consolidation relative to lower slope values (Kennedy et al., 2019).

[30] In the case of one endogenous regressor, the Kleibergen-Paap statistics is equivalent to the effective first-stage F statistic proposed by Montiel-Olea and Pflueger (2013).

in large-farm consolidation. This relationship is significant at 1% and implies that a one standard deviation increase in the terrain's inclination is associated with a decrease in large-farm consolidation of three percentage points. The reduced-form estimates are also statistically significant at 1% and are close in magnitude to the first-stage estimates, suggesting that the two patterns are closely related (Angrist and Pischke, 2008).

Table 1.3: Large-Farm Consolidation, Sectoral Employment & Unemployment Rate

Panel A. First Stage and Reduced Form Estimates			
	Share Area in Large Farms (Δ_{05-93}) (1)	Share Farm Employment (Δ_{05-93}) (2)	Unemployment Rate (Δ_{05-93}) (3)
Terrain's inclination (degrees)	-0.00239***	0.00186***	-0.00267***
	(0.000568)	(0.000656)	(0.000443)
Panel B. 2sls Estimates			
Share area in large farms (Δ_{05-93})		-0.777***	1.118***
		(0.301)	(00.309)
Standardized effect		-0.100	0.144
First Stage Results:			
Kleibergen-Paap F-statistic		17.69	17.69
Weak-Instrument Robust Inference:			
Anderson-Rubin P-value		0.004	0.000
Anderson-Rubin Confidence Set (95%)		[-1.59,-0.35]	[0.68,2.05]

Notes: 2sls estimates using Equation 1.1. The number of municipalities in these estimations is 590. All the specifications include province fixed effects and conflict-related covariates. Standard errors are clustered at the municipality level. The standardized effect is calculated by multiplying the coefficient of interest by one standard deviation of the change in large-farm consolidation (see Table 1.2). The null hypothesis of the Anderson-Rubin test is that the effect of land consolidation on the respective outcome is zero. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

The 2sls results indicate that rural economies with an increase in large-farm consolidation experienced a decline in the share of agricultural labor and a sizeable increase in unemployment. These estimates are significant at the 1% and remain unchanged when using the Anderson-Rubin inference procedure, suggesting that the results are not driven by a potential weak relationship in the first stage. To illustrate the magnitude of the estimates, consider the average Colombian municipality which in 1993 had a share of agricultural workers of 67% and an unemployment rate of 2.9%. If this municipality experienced a one standard deviation increase in consolidation –corresponding to a twelve percentage point rise in the share of area in large farms– , the agricultural employment share would fall ten percentage points, and the unemployment rate would increase 14 points.

I now explore how large-farm consolidation affected the wage of rural workers. Table 1.4 reports the two-sample 2sls results and provides support for the consistency of these estimates since I cannot reject the equality of the first-stage coefficient across the two samples used. The initial results suggest that the effect

on wages is negative and marginally significant at 10%. However, once I account for weak-instrument robust inference, the significance of the estimate vanishes, suggesting that large-farm consolidation did not have any significant effect on wages. Since these estimates do not properly control for income effects induced by the change in the terms of trade, it is important to regard them with caution. Additional results suggest that not accounting for such income trends may underestimate the impact on workers' wages (see Section 1.5.1).

Table 1.4: Large-Farm Consolidation and Log Hourly Wage (TS2sls)

	(1)	(2)
Share of area in large farms	-2.946	-3.249*
	(1.985)	(1.967)
Standardized effect	-0.370	-0.409
Control variables 1985	✓	✓
Conflict-related covariates		✓
First Stage Results:		
Kleibergen-Paap F-statistic	16.857	16.859
Test of equality of coefficients (pval)†	0.670	0.645
Weak-Instrument Robust Inference:		
Anderson-Rubin P-value	0.211	0.169
Anderson-Rubin Confidence Set (95%)	[-8.53,1.69]	[-8.70,1.26]

Notes: TS2sls estimates using Equation 1.2. The number of observations is 2,529. All the specifications include municipality fixed effects, department-year fixed effects, and interactions of both the population size and the share of rural population in 1985 with year fixed effects. Standard errors are robust to heteroskedasticity. The second row in *First Stage Results* tests the null hypothesis that the first-stage coefficient is equal across both samples, which is an additional assumption for the consistency of the two-sample 2sls estimator (†). The null hypothesis of the Anderson-Rubin test is that the effect of consolidation on the hourly wage is zero. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Taken together, the sizeable estimates on unemployment and the results on the wage suggest that the net effect of consolidation on workers' income is negative. My estimates on sectoral employment imply that land consolidation explains 72% of a standard deviation in the decline in farm employment between 1993 and 2005 across all municipalities.[31]

1.5.1 Robustness Exercises

Additional controls: As discussed in Section 1.4, one potential threat to the exclusion restriction is the upsurge in conflict during the period analysis. Historical accounts suggest that conflict and forced displacement had different intensities across municipalities in the lowlands and the slopes (Centro Nacional de Memoria Historica, 2013, 2012). In my preferred specification, I control for measures on the upsurge of conflict. However, as Table 1C shows, my main estimates are robust to the exclusion of these covariates, suggesting that municipalities of the same province had similar exposure to these shocks regardless of their topography. In

[31]This measure is calculated dividing the standardized effect of consolidation on the share of farmworkers (-0.10) by a standard deviation in the observed change of farmworkers share between 1993 and 2005 (0.14).

columns 3 and 6, I extend my preferred specification to control for potential trends driven by distinct levels of urbanization before the trade liberalization took place. While the estimates on unemployment decrease slightly, they remain significant at 1%.

Income effects induced by trade-liberalization: My preferred specification compares municipalities of the same province to account for the potential correlation between terrain's inclination -the instrument-, crop suitability, and income effects. In Table 1D, I show that province fixed effects play an important role in accounting for the expected bias generated by these income effects. The 2sls estimates that compare towns in the lowlands and the mountains without conditioning on province would have predicted a decline in unemployment rates after consolidation, since the economies in the lowlands experienced a general improvement in their terms of trade (Jaramillo, 2002).

This table also shows that province fixed effects do a better job in accounting for the expected bias on unemployment rates than department fixed effects. This suggests that my main findings on the effects of large-farm consolidation on wages may underestimate the real effect.

Alternative standard errors: In addition to inference procedures for weak instruments, the main results are robust to different types of intra-cluster correlation. As shown in Table 1E, the results are practically unchanged when using two-way clustered standard errors by municipality and province-year as well as municipality and department-year. The latter allows for correlation of the municipality's error across time and the error within-time correlation across municipalities of the same department.[32]

1.5.2 Potential Mechanisms

Overall, my findings indicate that consolidation led to a decrease in the demand for farmworkers and an absolute reduction in the demand for rural labor. These results are consistent with the framework's predictions when the pull response in the nonfarm sector is small relative to the push response in the farms. That is, the increase in local consumption, and thereby labor demand in the nonfarm sector, is not large enough to absorb the workers that are leaving the farms.

While the reallocation of labor across space is another mechanism through which local labor markets adjust in the long-term (Breza et al., 2021; Asher et al., 2021; Bustos et al., 2016), my results on unemployment suggest that migration was not large enough to offset the decrease in labor demand after consolidation. These findings are consistent with the low levels of integration across rural labor markets during the 1990s (Nupia, 1997) and support previous empirical results on the delayed adjustment of labor markets to economic shocks (Dix-Carneiro and Kovak, 2019; Dix-Carneiro, 2014; Autor et al., 2014; Artuç et al., 2010).

[32]To conduct this analysis, I estimate the panel fixed effects version of Equation 1.1 where the instrument is given by the interaction of the municipality's inclination terrain and the time trend, $Z_{mpt} = S_{mp} \times T_t$.

Due to data limitations, I cannot directly test the effect of land consolidation on local consumption to assess the importance of non-homothetic consumption growth as profits get concentrated. However, as an indirect analysis, I estimate Equation 1.1 for a subset of economies where local multiplier effects are expected to be large. In particular, towns that are most suitable to produce coffee, a crop that has historically been grown by producers who reside in their farms or the local economies (see Table 1F). This descriptive analysis indicates that the effect of consolidation on unemployment was smaller and less significant across family-oriented production regions, suggesting that local consumption is driving part of the absolute decrease in labor demand.

1.6 Conclusion

This paper provides empirical evidence of the effect of large-farm consolidation on rural labor markets in Colombia. The empirical findings indicate that consolidation can lead to are allocation of labor away from agriculture within rural economies along with a decrease in workers' earnings. These results imply a reduction in local labor demand after consolidation, and are consistent with a model that features differences in farm labor intensity and non-homothetic consumption growth in the nonfarm sector. My findings on workers' earnings imply that migration was not large enough to offset the decrease in labor demand over a period of twelve years. This suggests the existence of barriers to labor mobility across space and is in line with previous work on the delayed response of local labor markets to economy-wide shocks (Dix-Carneiro and Kovak, 2019; Dix-Carneiro, 2014).

This paper has several policy implications. First, it sheds light on the spatial impacts of structural transformation. Although this process is usually associated with increases in income at a national level, my results suggest that workers in rural areas may be negatively affected due to the aggregate decline in the demand for local labor and the limited migration over the medium term. My findings also shed light on the distributional impacts of land consolidation policies on rural welfare. I show that workers may experience a decrease in income despite the potential gains of farmers in agricultural productivity and profits (Adamopoulos and Restuccia, 2014; Adamopoulos et al., 2019).

While this paper is a step toward understanding the effect of large-farm consolidation on rural labor markets, it opens several questions for future research. For example, what are the main mechanisms? And how persistent are the effects beyond the medium term? Given the rise in consolidation policies in developing countries, it is also essential to assess the impacts of consolidation on the overall welfare of rural populations. This is an extension I am currently working on.

1.7 References

T. Adamopoulos and D. Restuccia. The size distribution of farms and international productivity differences. *The American Economic Review*, 104(6):1667–1697, 2014. ISSN 00028282. URL http://www.jstor.org/stable/42920862.

T. Adamopoulos, L. Brandt, J. Leight, and D. Restuccia. Misallocation, selection and productivity: A quantitative analysis with panel data from china misallocation, selection and productivity: A quantitative analysis with panel data from china †, 2019.

D. Ali, K. Deininger, and A. Harris. Does Large Farm Establishment Create Benefits for Neighboring Smallholders? Evidence from Ethiopia. *Land Economics*, 95(1):71–90, 2019. URL https://ideas.repec.org/a/uwp/landec/v95y2019i1p71-90.html.

I. Andrews, J. H. Stock, and L. Sun. Weak Instruments in Instrumental Variables Regression: Theory and Practice. *Annual Review of Economics*, 11, 2019. doi: 10.1146/annurev-economics. URL https://doi.org/10.1146/annurev-economics-.

J. D. Angrist and A. B. Krueger. The effect of age at school entry on educational attainment: An application of instrumental variables with moments from two samples. *Journal of the American Statistical Association*, 87:328–336, 1992. ISSN 1537274X. doi: 10.1080/01621459.1992.10475212. Goal: this is the paper that develops the two-sample IV estimator.

J. D. Angrist and A. B. Krueger. Split-sample instrumental variables estimates off the return to schooling. *Journal of Business and Economic Statistics*, 13:225–235, 1995. ISSN 15372707. doi: 10.2307/1392377. Goal: this paper derives assumptions and properties of the split-sample IV estimator. I also has some applications.

J. D. Angrist and J.-S. Pischke. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press, Dec. 2008. ISBN 0691120358.

M. A. Arias, A. M. Ibáñez, and A. Zambrano. Agricultural production amid conflict: Separating the effects of conflict into shocks and uncertainty. *World Development*, 2018. ISSN 18735991. doi: 10.1016/j.worlddev.2017.11.011.

J. Arteaga, C. C. Osorio, D. Cuéllar, A. M. Ibañez, R. Londoño Botero, M. Murcia, J. Neva, Á. Nieto, D. I. Rey, and F. Sánchez. Fondo de Tierras del Acuerdo Agrario de la Habana: Estimaciones y Propuestas Alternativas. Documentos CEDE 015630, Universidad de los Andes - CEDE, June 2017. URL https://ideas.repec.org/p/col/000089/015630.html.

E. Artuç, S. Chaudhuri, and J. McLaren. Trade shocks and labor adjustment: A structural empirical approach. *American Economic Review*, 100(3):1008–45, June 2010. doi: 10.1257/aer.100.3.1008. URL https://www.aeaweb.org/articles?id=10.1257/aer.100.3.1008.

S. Asher, A. Campion, D. Gollin, and P. Novosad. The long-run development impacts of agricultural productivity gains: Evidence from irrigation canals in india. Workingpaper, July 2021.

D. H. Autor, D. Dorn, G. H. Hanson, and J. Song. Trade Adjustment: Worker-Level Evidence. *The Quarterly Journal of Economics*, 129(4):1799–1860, 2014.

Á. Balcázar. Transformaciones en la agricultura colombiana entre 1990 y 2002. *Revista de Economía Institucional*, 5(9):128–145, 2003. ISSN 0124-5996.

C. B. Barrett, L. Christiaensen, M. Sheahan, and A. Shimeles. On the structural transformation of rural africa. *Journal of African Economies*, 26:11–35, 2017. doi: 10.1093/jae/ejx009. URL http://wdi.worldbank.org/.

P. Bottazzi, D. Crespo, L. O. Bangura, and S. Rist. Evaluating the livelihood impacts of a large-scale agricultural investment: Lessons from the case of a biofuel production company in northern Sierra Leone. *Land Use Policy*, 73(C):128–137, 2018. doi: 10.1016/j.landusepol.2017. URL https://ideas.repec.org/a/eee/lauspo/v73y2018icp128-137.html.

E. Breza, S. Kaur, and Y. Shamdasani. Labor rationing. *American Economic Review*, 111(10):3184–3224, October 2021. doi: 10.1257/aer.20201385. URL https://www.aeaweb.org/articles?id=10.1257/aer.20201385.

G. Bryan, S. Chowdhury, and A. M. Mobarak. Underinvestment in a profitable technology: The case of seasonal migration in bangladesh. *Econometrica*, 82(5):1671–1748, 2014. doi: https://doi.org/10.3982/ECTA10489. URL https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA10489.

P. Bustos, B. Caprettini, and J. Ponticelli. Agricultural productivity and structural transformation: Evidence from brazil. *American Economic Review*, 106:1320–1365, 2016. doi: 10.1257/aer.20131061. URL http://dx.doi.org/10.1257/aer.20131061.

P. Bustos, G. Garber, and J. Ponticelli. Capital Accumulation and Structural Transformation*. *The Quarterly Journal of Economics*, 135(2):1037–1094, 01 2020. ISSN 0033-5533. doi: 10.1093/qje/qjz044. URL https://doi.org/10.1093/qje/qjz044.

M. Carter. Identification of the Inverse Relationship between Farm Size and Productivity : An Empirical Analysis of Peasant Agricultural Production. *Oxford Economic Papers*, 36(1):131–145, 1984.

F. Caselli and W. J. Coleman II. The u.s. structural transformation and regional convergence: A reinterpretation. *Journal of Political Economy*, 109(3):584–616, 2001. doi: 10.1086/321015. URL https://doi.org/10.1086/321015.

Centro Nacional de Memoria Historica. *Justicia y paz. Tierras y territorios en las versiones de los paramilitares*. 2012. ISBN 9789896540821.

Centro Nacional de Memoria Historica. *¡Basta Ya! Colombia: Memorias de Guerra y Dignidad*. 2013. ISBN 9789585760844. URL www.centrodememoriahistorica.gov.co.

J. Choi, J. Gu, and S. Shen. Weak-instrument robust inference for two-sample instrumental variables regression. *Journal of Applied Econometrics*, 33(1):109–125, 2018. doi: 10.1002/jae.2580. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/jae.2580.

C. Clark. *The Conditions of Economic Progress*. Macmillan, 1940.

P. Collier and S. Dercon. African Agriculture in 50 Years: Smallholders in a Rapidly Changing World? *World Development*, 63:92–101, 2014. doi: 10.1016/j.worlddev.2013.10.001. URL http://dx.doi.org/10.1016/j.worlddev.2013.10.001.

DANE. Propuesta de Codificacion de Nuevas Divisiones Administrativas. Technical report, 2014. URL http://www.dane.gov.co/candane/.

C. A. Davis. Why Did Sugarcane Growers Suddenly Adopt Existing Technology? Working papers, 2017.

K. Deininger and F. Xia. Quantifying spillover effects from large land-based investment: The case of mozambique. *World Development*, 87(C):227–241, 2016. URL https://EconPapers.repec.org/RePEc:eee:wdevel:v:87:y:2016:i:c:p:227-241.

K. Deininger, D. Byerlee, J. Lindsay, A. Norton, H. Selod, and M. Stickler. Rising Global Interest in Farmland: Can It Yield Sustainable and Equitable Benefits? Technical report, 2011.

K. Deininger, S. Jin, Y. Liu, and S. K. Singh. Can labor market imperfections explain changes in the inverse farm size-productivity relationship? longitudinal evidence from rural india, 2016. URL http://econ.worldbank.org.

Departamendo Nacional de Planeación. Manual Metodológico para La Determinación de La Unidad Agricola Familiar Promedio Municipal. Technical report, 2000. URL https://www.dnp.gov.co/Portals/0/archivos/documentos/DDRS/Publicaciones{_}Estudios/ManualUAF.pdf.

K. Desmet and E. Rossi-Hansberg. Spatial development. *American Economic Review*, 104(4):1211–43, April 2014. doi: 10.1257/aer.104.4.1211. URL https://www.aeaweb.org/articles?id=10.1257/aer.104.4.1211.

R. Dix-Carneiro. Trade liberalization and labor market dynamics. *Econometrica*, 82(3):825–885, 2014. doi: https://doi.org/10.3982/ECTA10457. URL https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA10457.

R. Dix-Carneiro and B. K. Kovak. Trade liberalization and regional dynamics. *American Economic Review*, 107(10):2908–46, October 2017. doi: 10.1257/aer.20161214. URL https://www.aeaweb.org/articles?id=10.1257/aer.20161214.

R. Dix-Carneiro and B. K. Kovak. Margins of labor market adjustment to trade. *Journal of International Economics*, 117:125–142, 2019. ISSN 0022-1996. doi: https://doi.org/10.1016/j.jinteco.2019.01.005. URL https://www.sciencedirect.com/science/article/pii/S0022199619300078.

F. Eckert and M. Peters. Spatial structural change. Meeting Papers 98, 2018 Society for Economic Dynamics, 2018. URL https://ideas.repec.org/p/red/sed018/98.html.

M. Eswaran and A. Kotwal. Access to Capital and Agrarian Production Organisation. *The Economic Journal*, 96(382):482–498, 1986.

P. Fajgelbaum and S. J. Redding. Trade, Structural Transformation and Development: Evidence from Argentina 1869-1914. NBER Working Papers 20217, National Bureau of Economic Research, Inc, June 2018. URL https://ideas.repec.org/p/nbr/nberwo/20217.html.

F. Farrokhi and H. S. Pellegrina. Global trade and margins of productivity in agriculture. Working Paper 27350, National Bureau of Economic Research, June 2020.

M. Fernández, A. M. Ibáñez, and X. Peña. Adjusting the Labour Supply to Mitigate Violent Shocks: Evidence from Rural Colombia. *Journal of Development Studies*, 50(8):1135–1155, 2014. ISSN 17439140. doi: 10.1080/00220388.2014.919384. URL http://dx.doi.org/10.1080/00220388.2014.919384.

A. Foster and M. Rosenzweig. Economic Development and The Decline of Agricultural Employment. In *Handbook of Development Economics*, volume 4, pages Chapter–47. 2008.

A. D. Foster. Creating Good Employment Opportunities for the Rural Sector. 2011. URL https://www.adb.org/sites/default/files/publication/29119/economics-wp271.pdf.

A. D. Foster and M. R. Rosenzweig. Are There too Many Farms in the World? Labor-Market Transaction Costs, Machine Capacities and Optimal Farm Size. 2017.

J. García. Evolución de la distribución de las fincas cafeteras Hacía una regionalización de la caficultura colombiana. *Ensayos sobre Economía Cafetera, Federación Nacional de Cafeteros*, 19:193–213, 2003. URL https://www.federaciondecafeteros.org/static/files/3.evolucionfincascafeteras.pdf.

S. Haggblade, P. Hazell, and T. Reardon. *Transforming the Rural Nonfarm Economy: Opportunities and Threats in the Developing World*. International Food Policy Research Institute Series. Johns Hopkins University Press, 2007. ISBN 9780801886645. URL https://books.google.com/books?id=5QNHAwAAQBAJ.

S. Haggblade, P. B. R. Hazell, and T. Reardon. Transforming the rural nonfarm economy: Opportunities and threats in the developing world. Technical report, 2009.

J. Hamory, M. Kleemans, N. Y. Li, and E. Miguel. Reevaluating Agricultural Productivity Gaps with Longitudinal Microdata. *Journal of the European Economic Association*, 19(3):1522–1555, 11 2020. ISSN 1542-4766. doi: 10.1093/jeea/jvaa043. URL https://doi.org/10.1093/jeea/jvaa043.

B. Herrendorf, R. Rogerson, and Ákos Valentinyi. Chapter 6 - growth and structural transformation. In P. Aghion and S. N. Durlauf, editors, *Handbook of Economic Growth*, volume 2 of *Handbook of Economic Growth*, pages 855–941. Elsevier, 2014. doi: https://doi.org/10.1016/B978-0-444-53540-5.00006-9. URL https://www.sciencedirect.com/science/article/pii/B9780444535405000069.

R. T. Herrmann. Large-scale agricultural investments and smallholder welfare: A comparison of wage labor and outgrower channels in tanzania. *World Development*, 90:294–310, 2017. ISSN 0305-750X. doi: https://doi.org/10.1016/j.worlddev.2016.10.007. URL https://www.sciencedirect.com/science/article/pii/S0305750X16300390.

R. Hornbeck and S. Naidu. When the levee breaks: Black migration and economic development in the american south. *American Economic Review*, 104(3):963–90, March 2014. doi: 10.1257/aer.104.3.963. URL https://www.aeaweb.org/articles?id=10.1257/aer.104.3.963.

I. G. A. C. IGAC. Resolucion 2555 de 1988, 1988.

I. G. A. C. IGAC. *Atlas de la Distribucion de la Propiedad Rural en Colombia*. 2012.

A. Inoue and G. Solon. Two-sample instrumental variables estimators. *The Review of Economics and Statistics*, 92:557–561, 2010.

C. Jaramillo and A. M. Ibáñez. El censo nacional de población: una comparación de metodologías mediante simulaciones de Monte Carlo. *Coyuntura Social*, (32):43–64, 2005.

C. F. Jaramillo. La agricultura colombiana en la década del noventa. *Revista de Economía del Rosario*, 1998. ISSN 2145-454X.

C. F. Jaramillo. *Crisis y transformacion de la agricultura colombiana, 1990-2000*. Banco de la Republica, 2002. ISBN 9583800864.

X. Jiao, C. Smith-Hall, and I. Theilade. Rural household incomes and land grabbing in cambodia. *Land Use Policy*, 48:317–328, 2015. ISSN 0264-8377. doi: https://doi.org/10.1016/j.landusepol.2015.06.008. URL https://www.sciencedirect.com/science/article/pii/S0264837715001763.

E. H. Kennedy, S. Lorch, and D. S. Small. Robust causal inference with continuous instruments using the local instrumental variable curve. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81(1):121–143, 2019. doi: 10.1111/rssb.12300. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/rssb.12300.

S. Kuznets. Quantitative aspects of the economic growth of nations: Ii. industrial distribution of national product and labor force. *Economic Development and Cultural Change*, 5(4):1–111, 1957. ISSN 00130079, 15392988. URL http://www.jstor.org/stable/1151943.

J. Lanjouw and P. Lanjouw. The rural non-farm sector: issues and evidence from developing countries. *Agricultural Economics*, 26(1):1–23, 2001. URL https://EconPapers.repec.org/RePEc:eee:agecon:v:26:y:2001:i:1:p:1-23.

A. A. Laskievic. Commodity booms and structural transformation: the role of input use and land inequality. Workingpaper, July 2021.

G. Le Roux. Comparabilidad de los censos colombianos de 1993 y 2005: cambios en la recolección de información y dificultades en el análisis de las evoluciones intraurbanas en Bogotá. *Cuadernos de Vivienda y Urbanismo*, 2013.

W. A. Lewis. Economic development with unlimited supplies of labour. *The Manchester School of Economic and Social Studies*, 22:139–191, 1954. ISSN 0025-2034. doi: 10.1111/j.1467-9957.1954.tb00021.x.

C. Liao, S. Jung, D. G. Brown, and A. Agrawal. Spatial patterns of large-scale land transactions and their potential socio-environmental outcomes in cambodia, ethiopia, liberia, and peru. *Land Degradation & Development*, 31(10):1241–1251, 2020. doi: https://doi.org/10.1002/ldr.3544. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/ldr.3544.

S. K. Lowder, J. Skoet, and T. Raney. The number, size, and distribution of farms, smallholder farms, and family farms worldwide. *World Development*, 87:16 – 29, 2016. ISSN 0305-750X. doi: https://doi.org/10.1016/j.worlddev.2015.10.041. URL http://www.sciencedirect.com/science/article/pii/S0305750X15002703.

A. Machado and R. Suarez. *El mercado de tierras en Colombia : una alternativa viable?* Tercer Mundo Editores : CEGA : IICA, Santafé de Bogotá, Colombia, 1999. ISBN 958601858X.

C. U. Mallarino. Borrón y cuenta nueva: las estadísticas en Colombia se reinventan a sí mismas. *Universitas Humanistica*, 63(63), 2007.

L. R. Martinez. Sources of revenue and government performance: Evidence from colombia. Workingpaper, Oct. 2020.

J. W. McArthur and G. C. McCord. Fertilizing growth: Agricultural inputs and their effects in economic development. *Journal of Development Economics*, 127:133–152, 2017. ISSN 0304-3878. doi: https://doi.org/10.1016/j.jdeveco.2017.02.007. URL https://www.sciencedirect.com/science/article/pii/S0304387817300172.

E. B. McCullough. Labor productivity and employment gaps in sub-saharan africa. *Food Policy*, 67:133–152, 2017. ISSN 0306-9192. doi: https://doi.org/10.1016/j.foodpol.2016.09.013. URL https://www.sciencedirect.com/science/article/pii/S0306919216303803. Agriculture in Africa – Telling Myths from Facts.

M. Mcmillan, D. Rodrik, and C. Sepúlveda. *Structural Change, Fundamentals, and Growth: A Framework and Case Studies*. International Food Policy Research Institute (IFPRI), 2017. doi: 10.2499/9780896292147. URL http://drodrik.scholar.harvard.edu/files/dani-rodrik/files/structural_change_fundamentals_and_growth.pdf.

J. W. Mellor. *Agricultural Development and Economic Transformation Promoting Growth with Poverty Reduction*. Palgrave Macmillan, 2017. URL http://www.springer.com/series/14651.

G. Michaels, F. Rauch, and S. J. Redding. Urbanization and Structural Transformation *. *The Quarterly Journal of Economics*, 127(2):535–586, 03 2012. ISSN 0033-5533. doi: 10.1093/qje/qjs003. URL https://doi.org/10.1093/qje/qjs003.

J. L. Montiel-Olea and C. Pflueger. A robust test for weak instruments. *Journal of Business & Economic Statistics*, 31(3):358–369, 2013. doi: 10.1080/00401706.2013.806694. URL https://doi.org/10.1080/00401706.2013.806694.

M. Morten and J. Oliveira. The Effects of Roads on Trade and Migration: Evidence from a Planned Capital City. Technical report, 2018.

D. Nagy. City location and economic development. Technical report, 2016.

O. A. Nupia. Integración espacial en los mercados laborales: evidencia para las regiones colombianas. *Desarrollo y Sociedad*, (40), 1997.

J. A. Ocampo. Marco Conceptual de la Misión para la Transformación del Campo. Technical report, 2014.

K. Ortiz-Becerra. Large-farm consolidation and welfare in rural economies, 2022.

D. Pacini and F. Windmeijer. Robust inference for the Two-Sample 2SLS estimator: Appendix. *Economics Letters*, pages 1–8, 2016.

J. A. Pinzón and J. Fonti. Una aproximación al catastro en Colombia. *Revista UD y La Geomática*, 1(1):25–46, 2007.

M. Ravallion and G. Datt. When Is Growth Pro-Poor? Evidence from the Diverse Experiences of India's States. Technical report, 1999.

T. Reardon, J. Berdegué, and G. Escobar. Rural nonfarm employment and incomes in latin america: Overview and policy implications. *World Development*, 29(3):395–409, 2001. ISSN 0305-750X. doi: https://doi.org/10.1016/S0305-750X(00)00112-1. URL https://www.sciencedirect.com/science/article/pii/S0305750X00001121. Rural Nonfarm Employment and Incomes in Latin America.

A. Sen. Market failure and control of labour power: towards an explanation of 'structure 9 and change in Indian agriculture. Part 1. *Cambridge Journal of Economics*, 5:201–228, 1981. URL https://academic.oup.com/cje/article-abstract/5/3/201/1686512.

M. Uribe Castro. Caffeinated development: Export sector, human capital, and structural transformation in colombia. Workingpaper, Feb. 2020.

1A Additional Tables

Table 1A: Pre-Trends Test for Analysis on Sectoral Employment & Unemployment Rate

	Share Rural Pop. (Δ_{85-93})		Total Population (Δ_{85-93})	
	(1)	(2)	(1)	(2)
Terrain's Inclination (degrees)	0.0001	0.0001	-31.39**	-20.62
	(0.0003)	(0.0004)	(14.46)	(14.57)
Province fixed effects	✓	✓	✓	✓
Conflict-related events (Δ_{85-93})		✓		✓
Number of municipalities	576	576	576	576

Notes: OLS estimates of the (Δ_{85-93}) in share of rural population and total population on the instrument before the observed change in land consolidation. These variables measure urbanization levels and are considered proxy variables for the main outcomes of interest. Standard errors are clustered at the municipality level. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 1B: Pre-Trends Test for Analysis on Log Hourly Wage

	Share Rural Population (Δ_{85-93})		Total Population (Δ_{85-93})	
	(1)	(2)	(1)	(2)
Terrain's Inclination (degrees)	0.00009	0.00001	-48.93***	-36.03***
	(0.00025)	(0.00024)	(13.97)	(10.41)
Department fixed effects	✓	✓	✓	✓
Conflict-related covariates (Δ_{85-93})		✓		✓
Number of municipalities	576	576	576	576

Notes: OLS estimates of the (Δ_{85-93}) in share of rural population and total population on the instrument before the observed change in land consolidation. These variables measure urbanization levels and are considered proxy variables for the main outcomes of interest. Standard errors are clustered at the municipality level. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 1C: Large-Farm Consolidation, Sectoral Employment & Unemployment Rate: Different Set of Covariates

	Share of Agricultural Emp. (Δ_{05-93})			Unemployment Rate (Δ_{05-93})		
	(1)	(2)	(3)	(1)	(2)	(3)
Share area large farms (Δ_{05-93})	-0.777***	-0.780***	-0.845***	1.118***	1.129***	0.856***
	(0.301)	(0.293)	(0.309)	(0.309)	(0.307)	(0.246)
Standardized effect	-0.100	-0.100	-0.109	0.144	0.145	0.110
Conflict-related events (Δ_{05-93})	✓		✓	✓		✓
Population in 1985 (logs)			✓			✓
First Stage Results:						
Kleibergen-Paap F-statistic	17.69	18.69	16.37	17.69	18.69	16.37
Weak-Instrument Robust Inference:						
Anderson-Rubin P-value	0.004	0.002	0.001	0.000	0.000	0.000
A-R Confidence Set (95%)	[-1.58, -0.35]	[-1.57, -0.36]	[-1.34, -0.11]	[0.68,2.05]	[0.69,2.05]	[0.51,1.67]
N	590	590	576	590	590	576

Notes: 2sls estimates using different covariates in Equation 1.1. Results in (1) refer to main specification. All specifications include province fixed effects and standard errors are clustered at the municipality level. The null hypothesis of the Anderson-Rubin test is that the effect of consolidation on the respective outcome is zero. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 1D: Large-Farm Consolidation, Sectoral Employment & Unemployment Rate: The Importance of Province Fixed Effects

	Share of Agricultural Emp. (Δ_{05-93})			Unemployment Rate (Δ_{05-93})		
	(1)	(2)	(3)	(1)	(2)	(3)
Share area large farms (Δ_{05-93})	1.198***	-0.591**	-0.777**	-0.650***	1.034***	1.118***
	(0.169)	(0.277)	(0.301)	(0.100)	(0.315)	(0.309)
Standardized effect	0.154	-0.076	-0.100	-0.084	0.133	0.144
Department fixed effects		✓			✓	
Province fixed effects			✓			✓
First Stage Results:						
Kleibergen-Paap F-statistic	87.61	17.65	17.69	87.61	17.65	17.69
Weak-Instrument Robust Inference:						
Anderson-Rubin P-value	0.000	0.024	0.004	0.000	0.000	0.000

Notes: 2sls estimates using different sets of fixed effects in Equation 1.1. Results in (3) refer to main specification. The number of municipalities in these estimations is 590. All the specifications include conflict-related covariates and standard errors are clustered at the municipality level. The null hypothesis of the Anderson-Rubin test is that the effect of consolidation on the respective outcome is zero. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 1E: Large-Farm Consolidation, Sectoral Employment, & Unemployment Rate: Different Types of Standard Errors

	Share of Agricultural Emp.			Unemployment Rate		
	(1)	(2)	(3)	(1)	(2)	(3)
Share of area in large farms	-0.777***	-0.777**	-0.777**	1.118***	1.118***	1.118***
	(0.301)	(0.347)	(0.341)	(0.309)	(0.405)	(0.423)
Standardized effect	-0.100	-0.100	-0.100	0.144	0.1441	0.144
Standard errors:						
Cluster municipality level	✓			✓		
Two-way cluster mun & depto-year		✓			✓	
Two-way cluster mun & prov-year			✓			✓
First Stage Results:						
Kleibergen-Paap F-statistic	17.69	12.19	9.06	17.69	12.19	9.06
Weak-Instrument Robust Inference:						
Anderson-Rubin *P*-value	0.004	0.014	0.010	0.000	0.001	0.000
A-R Confidence Set (95%)	[-1.58, -0.35]	[-2.04, -0.28]	[-2.24, -0.29]	[0.68,2.05]	[0.54,2.72]	[0.65,3.46]

Notes: 2sls estimates using a fixed effects version of Equation 1.1. Results in (1) refers to main specification. The number of municipalities in these estimations is 590. All the specifications include municipality fixed effects, province-year fixed effects and conflict-related covariates. The null hypothesis of the Anderson-Rubin test is that the effect of consolidation on the respective outcome is zero. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

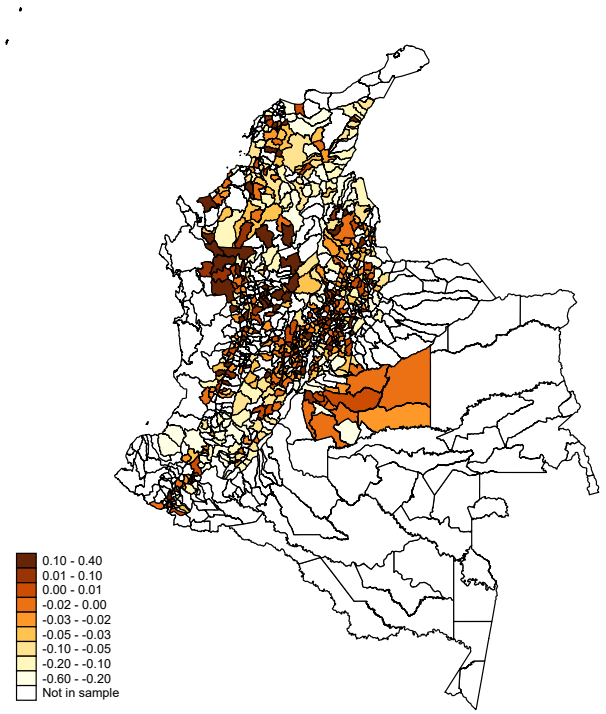
Table 1F: Large-Farm Consolidation and Unemployment Rate: By Expected Strength of Local Multiplier Effects

	Full Sample	Expected Local Mult. Effects	
		Strong	Weak
Share area large farms (Δ_{05-93})	1.118***	0.857*	1.041***
	(0.309)	(0.470)	(0.320)
Standardized effect	0.144	0.110	0.134
Average unemployment rate in 1993	2.5%	1.9%	2.7%
Kleibergen-Paap F-statistic	17.69	7.84	14.52
Weak-Inst. Robust Inference:			
Anderson-Rubin Conf. Set (95%)	[0.678, 2.049]	[0.187, 3.015]	[0.584, 2.11]
N	590	133	457

Notes: 2sls estimates using Equation 1.1. Outcome is (Δ_{05-93}) unemployment rate. The subsample with expected strong multiplier effects include economies between 1,200 and 1,800 m.a.s.l, which corresponds to towns with high suitability for coffee production. All specifications include province fixed effects and conflict-related covariates. Standard errors clustered at municipality level. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

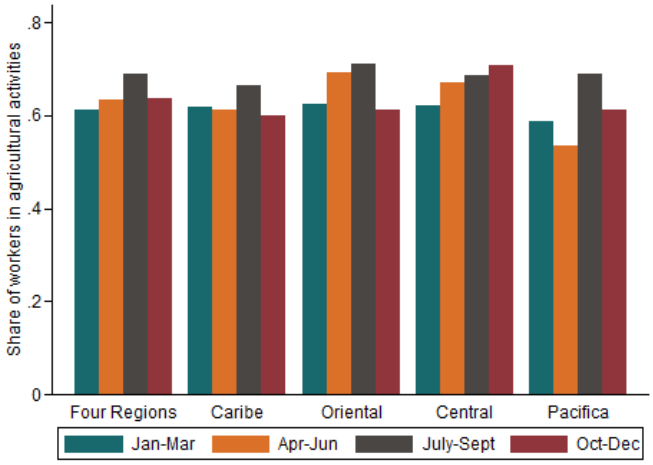
1B Additional Figures

Figure 1A: Spatial Variation in Consolidation Shift



Notes: This map displays the change in the share of area in large farms (in percentage points) for rural municipalities between 1993 and 2005. Own calculations using cadastral data.

Figure 1B: Share of Agricultural Employment in Rural Areas, by Quarter and Region



Notes: Own illustrations using data from the National Household Surveys.

ACKNOWLEDGMENTS

I am deeply grateful to Ashish Shenoy, Michael Carter, Dalia Ghanem, and Steve Boucher for providing guidance and advising on this project and throughout my degree. Many thanks to Douglas Gollin, Diego Restuccia, Jason Shogren, Gaurav Chiplunkar, Esteban Quiñonez, Ameet Morjaria, Debi Mohapatra, John Morehouse, Eleanor Wiseman, Ana Maria Ibanez, Fabio Sanchez, Juan Carlos Munoz, Margarita Gafaro, Jessica Rudder, Aleksander Michuda, and seminar audiences at MWIEDC, UC Davis, the 1st Workshop in Applied Microeconomics, and the LACEA-RIDGE Job Market Showcase for helpful discussions and valuable feedback. I also thank Olam International, its team of in-country operators, and the farmers in rural regions of Antioquia, Huila, Casanare, and Meta for their time during my fieldwork visits throughout the summer and fall of 2019. I acknowledge the support from the UC Davis Provost's Dissertation Fellowship, the Bert and Nell Krantz Fellowship, and the Henry A. Jastro Graduate Research Award.

Chapter 2

Large-Farm Consolidation and Welfare in Rural Economies

2.1 Introduction

The consolidation of agricultural land is accelerating in low and middle-income countries. In the past two decades, more than four thousand large-scale deals have taken place.[1] Close to 60% of these deals are transactions of at least five thousand hectares, and the median size across all transactions is seven thousand hectares. This growing interest in cropland has been primarily driven by the rise in the global demand for food and agricultural commodities, and it is unlikely to slow, given the current trends in population growth (Liao et al., 2020; Deininger et al., 2011). Thus, it is crucial to understand the implications of large-farm consolidation on the development and welfare of the rural economies where it takes place.

Land allocation has long been considered an important determinant of agricultural productivity and income (Chayanov, 1966; Sen, 1966; Berry and Cline, 1979). This relationship is driven by several countervailing forces that distort access to factor and output markets across producers with different farm sizes (Putterman, 1983; Feder, 1985; Eswaran and Kotwal, 1986; Carter and Kalfayan, 1989; Chavas, 2001; Collier and Dercon, 2014). On one hand, small producers have been found to be more labor efficient due to lower costs of hiring and labor supervision. Meanwhile, large producers often benefit from pecuniary economies and have better access to capital markets and modern value chains. Although these offsetting forces may vary across settings, recent empirical evidence suggests that large-scale production can lead to productivity gains. In particular, several studies in low and middle-income countries document that large producers are

[1] These numbers refer to completed transactions of at least five hundred hectares for agricultural activities. For more details on this data, see the Land Matrix Project.

often equally or more productive than farmers that operate at a small scale (Deininger et al., 2016; Helfand and Taylor, 2021; Foster and Rosenzweig, 2022). This observation is especially salient in settings where farmers can substitute labor for machinery, indicating that the cost advantages of small producers are likely to be outweighed by economies of scale in modern agriculture.[2]

Despite these potential gains of scale on agricultural productivity, however, the aggregate effect of large-farm consolidation on the welfare of rural populations is less clear. For one, limited access to credit markets can prevent smallholders from being competitive in land markets. Moreover, many rural households in developing countries are landless and derive most of their income from local labor markets. These labor markets are driven by the linkages between the agricultural and rural nonagricultural sectors. Thus, whether rural workers benefit from land consolidation in the short and medium-term will depend on local employment gains.

In this paper, I develop a quantitative model to study the effects of large-farm consolidation on the welfare of farmers and workers in rural economies. I focus on two main questions of interest. First, how does consolidation affect aggregate welfare? Second, who benefits when this consolidation takes place? Additionally, I examine how these welfare effects vary by type of consolidation. In particular, whether the rise of large operations is driven by merging the existing smallest plots or the merging of middle-sized farms.

I study these questions in the context of rural Colombia, where land concentration has been historically high, and 40% to 50% of agricultural land is consolidated in large farms. Two main features make this setting particularly relevant to study these questions. First, a majority of rural households are landless (58%) and derive most of their income from labor markets. Second, one of the provisions of the 2016 Peace Agreement is to redistribute at least 2.6 million hectares to landless peasants and farmers with insufficient acreage (Arteaga et al., 2017). This analysis is also relevant for other developing contexts where large-scale land deals are taking place since many of the stylized patterns that I document for Colombia are likely general to rural economies in low and middle-income countries.

I build a two-sector model of a rural economy featuring heterogeneity in the farm size distribution and production in agricultural and nonagricultural activities. The novelty of the model lies in combining several insights from the literature to shed light on the distributional impacts of consolidation across producers and workers. First, large-scale production is less labor intensive and more productive due to the substitution of labor for machinery. Furthermore, profit concentration influences the aggregate demand for rural workers since the expenditure share in the non-tradables produced in the nonfarm sector varies with income (Ranis and Stewart, 1993; Mellor, 2017). A key feature of this framework is that aggregate farm profits increase

[2]Of course, some crops can still be produced competitively at different scales depending on local factor endowments and labor costs (Deininger et al., 2011). Specialty and certified coffee, for instance, are often more profitable when produced at a small scale.

with the scale of agrarian operation, while the income of rural workers depend on the employment effects of land consolidation.

I show that these employment effects are determined by the interplay of two opposing responses in labor demand across both economic sectors. In particular, consolidation reduces the aggregate demand for workers if the pull response in the nonfarm sector is not large enough to offset the push of labor on the farms. In the short term, this decline in the demand for labor leads to a reduction in the income of rural workers since the low mobility across locations limits the adjustment of labor supply. Thus, the overall impact of consolidation on aggregate welfare depends on whether the positive productivity effects offset the potential losses in terms of employment.

I calibrate a baseline economy to farm-level and aggregate observations in Colombia. In particular, I approximate the farm size distribution with a Pareto density and choose its shape parameters to match a Gini index of 0.86, as estimated in previous studies (IGAC, 2012). To capture the decline in labor intensity across farm sizes, I match the labor-to-land ratio between the largest and smallest farms. In addition, I use data on household expenditures to identify the parameters of the non-homothetic preferences. The calibrated model resembles the distribution of farms and area implied by the data from the agricultural census. It also explains the direction of the farm size and input intensity relationship. This model, however, currently underestimates the decline in labor intensity. Thus, while I show that it has a meaningful power in explaining previous impacts of consolidation on rural employment, this quantitative framework should be regarded as a work in progress.[3]

Equipped with this quantitative framework, I first conduct a counterfactual analysis to examine the impact of large-farm consolidation on the structure of rural employment and wages. In particular, I assume that the land distribution remains Pareto and increase the proportion of area in large farms by thirteen percentage points. This counterfactual emulates one standard deviation of the observed change in consolidation across rural counties over the nineties, and provides a helpful benchmark to assess the explanatory power of the model. This analysis shows that consolidation leads to a four percentage points decrease in the proportion of farmworkers and an 8% decline in the wage of rural economies, implying that the structural transformation of rural economies - induced by the increase in large-farm operations - is accompanied by a decrease in workers' income.

These findings are in consonance with existing quasi-experimental evidence on the impacts of large-farm consolidation on the structural transformation of rural Colombia during the 90s (Ortiz-Becerra, 2022). They are qualitatively consistent with the estimated effects on workers' income and explain 40% of the structural

[3] A future version of this draft will consider a variation of the model that can reproduce the quantitative relationship between farm size and labor intensity.

change induced by consolidation during this period.[4] These results imply that my model has a relevant quantitative power in explaining the medium-term impacts of consolidation despite abstracting from spatial labor mobility, suggesting the existence of barriers to migration in rural economies beyond the short term.

Next, I examine the impact of land consolidation on the welfare of rural populations. A key finding from this analysis is that consolidation reduces rural welfare despite increasing agricultural income. This expansion of large-farm operations increases aggregate profits by 7% at the expense of a reduction in social welfare of more than 50%. These welfare effects, however, vary substantially across workers and producers. On the one hand, farmers benefit from increased agricultural productivity and experience an average welfare growth of 54%. In contrast, workers are adversely affected (14% decline in average welfare) by an overall decrease in rural labor demand. These findings suggest that large-farm consolidation has heterogeneous impacts across groups of individuals and that boosting aggregate output is insufficient to generate broad-based rural income growth.

Finally, I conduct two (counterfactual) redistribution policies to examine whether the effects of large-farm consolidation vary by the type of consolidation that takes place. The model suggests that the strength of the push and pull responses in the farm and nonfarm sectors differ across the farm size distribution. The first policy allocates multiple small plots into one large farm operated by a single producer, generating a substantial rise in the share of landless in the economy. In contrast, the second policy creates a large farm by combining land from several middle-sized plots. Although both policies lead to a decline in the proportion of farmworkers, there are some important differences in the impact of these policies on aggregate outcomes. In particular, merging smaller farms leads to a steeper decline in the wage of rural workers despite the more substantial gains in agricultural income and productivity. This consolidation policy also exacerbates the fall in aggregate social welfare relative to the scenario in which the land of middle-sized farms is combined (-34% vs. -12%), suggesting that the impacts on welfare depend on the size of the farms that are merged in order to form large-scale operations. These quantified effects, however, may underestimate the real difference in the impacts of both types of consolidation on welfare, as the model abstracts from the potential cost advantages of producing at a small scale.[5]

This paper contributes to the long-standing literature that examines the economic impacts of land allocation on economic development. This literature includes theoretical work that micro-founds the connections between scale and aggregate productivity (Carter and Kalfayan, 1989; Eswaran and Kotwal, 1986; Ma et al., 2021; Foster and Rosenzweig, 2022), and macroeconomic models that quantify the contribution of misal-

[4] See Ortiz-Becerra (2022) for more details on the medium-term effects of large-farm consolidation on rural labor markets in Colombia during the 1990s.

[5] As shown in Foster and Rosenzweig (2022), this additional countervailing force could lead to a u-shaped relationship between farm size and agricultural productivity, as opposed to a direct positive relation between productivity and scale.

location to productivity gaps (Adamopoulos and Restuccia, 2014; Chen, 2017; Adamopoulos et al., 2019; Adamopoulos and Restuccia, 2020; Santaeulàlia-Llopis, 2021).[6] My work complements these studies by examining implications for local labor markets and adds to this literature by introducing a framework that considers the links between the scale of production and labor demand in both the farm and nonfarm sectors. This framework combines insights from the relationship between farm size and productivity with two insights from the literature on the rural nonfarm economy: the importance of local consumption on employment in the nonfarm sector (Foster, 2011; Haggblade et al., 2009, 2007; Foster and Rosenzweig, 2008; Lanjouw and Lanjouw, 2001), and the connection between income concentration and labor demand (Mellor, 2017; Ranis and Stewart, 1993). By embedding these insights into a general equilibrium model, I show that large-farm consolidation has different impacts across farmers and workers, since it leads to a decline in the aggregate demand for labor despite the productivity gains.

This paper also contributes to the growing empirical work that examines the effects of large-scale transactions on the welfare of rural populations (Liao et al., 2020; Ali et al., 2019; Deininger and Xia, 2016). Most of these studies focus on studying spillover effects on smallholders' investments and income. My work quantifies economy-wide impacts on employment, productivity, and income, as well as distributional effects across farmers and workers. In addition, the framework developed in this paper can be easily calibrated using data from standard sources to quantify the impacts of land consolidation in other settings.

A second line of work related to this paper is the literature on structural transformation (Herrendorf et al., 2014). My findings show that land consolidation is another driver of the reallocation of labor across sectors. These findings also indicate that the reallocation of labor out of the farm sector can occur along with a decrease in income, which is the opposite relationship than we usually observe with the process of structural transformation. These results are related to recent studies that examine the implications of structural transformation across space (Bustos et al., 2020; Eckert and Peters, 2018; Nagy, 2016; Desmet and Rossi-Hansberg, 2014; Michaels et al., 2012; Caselli and Coleman II, 2001).

This article proceeds as follows. The next section provides an overview of the data and context. In section 2.3, I establish empirical patterns pertaining to farm size and employment in each sector. Section 2.4 describes the model, and section 2.5 presents the calibration of the baseline economy to Colombian data. Finally, section 2.6 examines the quantitative effects of large-farm consolidation on rural employment and income, and section 2.7 concludes.

[6]There is a wide strand of articles that complement this literature by examining the political economy effects of land inequality on economic development. See, for instance, Faguet et al. (2020) for a recent analysis in the Colombian setting.

2.2 Context and Data

2.2.1 Data

The primary data source for this analysis is the 2014 National Agricultural Census. I use this data to document empirical patterns, calibrate several of the parameters in the model, and assess the model's ability to match non-targeted dimensions at the farm-level.

The National Agricultural Census is an instrument that collects information on agricultural production and practices in the country. This census covers the scattered rural area from all the municipalities and includes all the agricultural production units (UPAs) regardless of titling and tenure regime.[7] From the 2.4 million agricultural units that are part of this census, 86% are held by private individuals and companies, and 14% are held by communities under collective rights. For each production unit, this instrument collects data on location, size, land use, number of workers, machinery possession, credit access, and livestock inventory. In addition, it collects information on production, final destination, and agricultural practices for each crop produced. This survey also contains data on the number of households that occupy the UPA, the dwellings' features, and the sociodemographic characteristics of the habitual residents.

In this analysis, I focus on the set of agricultural production units held by private individuals and companies. These units have an average area size of 17.2 hectares and account for seventy-four percent of the total farmland. The main variables that I use in this analysis are the number of workers, machinery possession, type of produced crops, and area size. Section 2A in the appendix provides details on these variables and the sample selection criteria.

In addition to this census, I also use the following five sources to document empirical patterns and calculate the moments for the model's calibration: National Household Surveys, the National Households' Budget Survey,and data on agricultural production costs from the study conducted by Perfetti et al. (2012). In this analysis, I focus on the data collected before 2020 to avoid deviations from regular patterns due to the economic impacts of the COVID-19 pandemic. See Section 2A in the appendix for details about these sources and the definition of variables.

2.2.2 Agriculture and Land

Agriculture is an important sector in Colombia's economy. It accounts for 7% of the gross domestic product and 19% of the value of exports. This sector also employs close to 22% of all workers in the country. In

[7] An UPA is defined as the set of plots under sole management that uses all or part of its area for agricultural purposes (DANE, 2016). It can be composed of one or more rural properties as long as they share at least one of the following production inputs: labor, machinery, equipment, or facilities.

terms of area, the land in agricultural use amounts to 42.3 million hectares, an area that is similar to the size of countries like Germany or Spain. From these, 80% are used to raise livestock and 20% to produce crops.

Due to its location in the tropics and the prevalence of mountain ranges, the country counts on a diversity of climates that are conducive to produce a wide variety of crops throughout the year. The main crops include seasonals such as rice, corn and yucca as well as perennials such as coffee, plantains, and oil palm (see Table 2A in the appendix). In terms of livestock, the most prominent animals raised are poultry, cattle, and porcine. Extensive cattle ranching is prevalent in the lowlands and inter-Andean valleys since terrains are flatter and the carrying capacity of the land is high. Of all the farms, 99% are operated by a natural person, and 1% are managed by firms. Most farmers sell their products in local and external markets, and only a tiny minority (2%) use them exclusively for their own consumption.

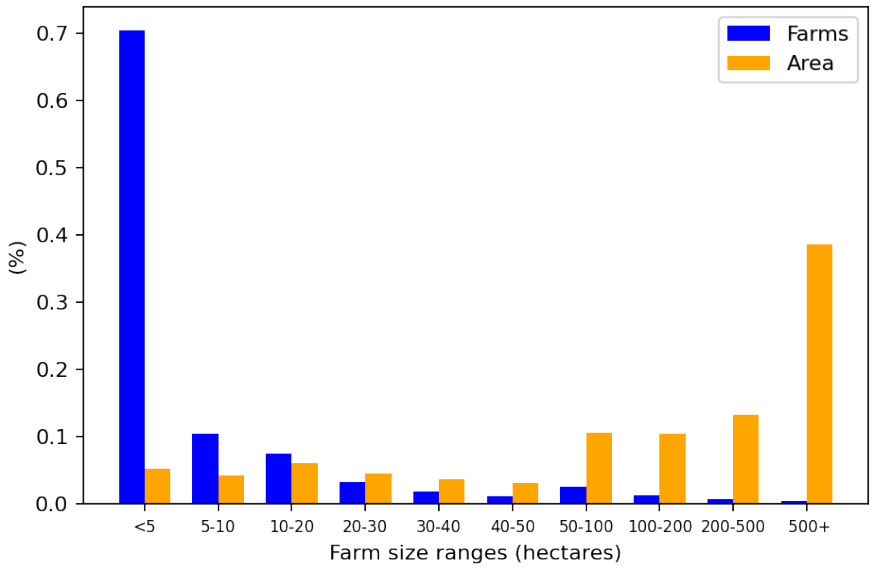
Figure 2.1 presents the distribution of farms and area across ten different size ranges. Farms below fifty hectares are considered small while farms with five hundred or more hectares are considered large.[8] The main patterns that emerge are the wide variation in the scale of agricultural production and the consolidation of land in large farms. Smallholders account for 95% of farms and only 27% of total area. In contrast, large farms only represent 0.4% of all production units and account for 40% of the land. Some of the determinants of this consolidation include the uneven distribution during colonial rule, the disorderly frontier expansion process, and the ineffective allocation of public lands throughout the twentieth century (Gáfaro et al., 2014). More recently, large-farm consolidation has been driven by different market forces that render scale production more profitable and the vast number of dispossessions that occurred during the upsurge of conflict in the past decades.[9]

Finally, one important feature of the land context is the high prevalence of individuals with informal property rights. Close to 48% of all private holders do not have a formal title to their land (Arteaga et al., 2017). This informality in property rights has led to a limited role of rental markets that persists over time. According to a report conducted by the National Department of Statistics, less than ten percent of producers operate rented land and this proportion has only grown four percentage points in the last four decades (DANE, 2015).

[8]This classification is based on a policy instrument that indicates the minimum plot size to generate income surplus given the agro-ecological features of the land (Departamendo Nacional de Planeación, 2000). This minimum plot size is known as the family farm unit and varies across subnational levels. The weighted average family unit across all municipalities is close to fifty hectares. Previous studies have used a threshold of ten family units (500 hectares) to define a large farm (Faguet et al., 2020; Machado and Suarez, 1999).

[9]See Ortiz-Becerra (2022), Fajardo (2014), and Balcázar (2003) for more details on land consolidation after the trade liberalization. For further information on land grabbing and dispossession during the civil conflict, see Centro Nacional de Memoria Historica (2012) and Centro Nacional de Memoria Historica (2013).

Figure 2.1: Distribution of Farms and Area



Notes: This figure presents the distribution of farms and area across different size ranges using data from the 2014 National Agricultural Census. See Section 2A for details on data sources and definitions.

2.2.3 Rural Labor Markets

Rural labor markets play a key role in the country's economy. They account for one-fifth of the total labor force and employment and are the main source of income for the bulk of rural households without access to land (60%). Relative to the cities, these economies are characterized by having a higher incidence of self-employment and informality. Therefore, while unemployment is low (7%), the average labor income for the majority of workers is below the minimum wage (Otero-Cortés, 2019).[10]

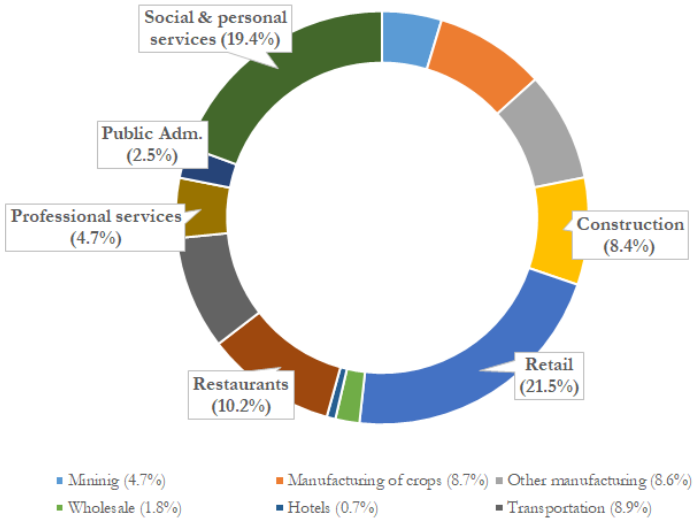
Overall, rural employment is driven by two main sectors of the economy. The first sector is characterized by the intensive use of land to raise livestock and produce crops (i.e. agriculture), and the second one is comprised of all activities that take place in other industries such as manufacturing and retail. Of the 5 million workers, sixty percent work in agriculture and forty percent work in non-farm activities. This employment structure is fairly stable throughout the year since the two rainy seasons in most parts of the country lead to a relatively steady demand for agricultural workers (see Figure 2A in the appendix).

One distinct feature of rural labor markets, relative to the urban sector, is that a large majority of the workers in the nonfarm sector (70%) work in industries that produce nontradable services or goods with high local value-added. For instance, 19% work in the provision of personal and social services, 8% work on

[10]Rural labor markets are more prominent in size when considering the new definition of rurality proposed by the *2014 Mission for the Transformation of the Countryside*. In this case, the proportion of rural workers amounts to 33% of total workers in the country. For a detailed characterization of rural labor markets using this new definition of rurality, see Tenjo and Jaimes (2015).

construction, and 22% work in retail.[11] Conversely, transportation employs 9% of the workers, and tradable industries such as food processing account for less than 10% of nonfarm jobs.[12] This observation implies that local consumption is a crucial driver of employment in the nonfarm sector, and thus, the aggregate demand for rural workers depends on how much of the aggregate income is spent in the local economies.

Figure 2.2: Distribution of Nonfarm Employment



Notes: This figure presents the distribution of rural nonfarm employment by type of industry using data from the national houseold surveys. See Section 2A for details on data sources and definitions.

2.3 Empirical Patterns Relating Farm Size and Labor Demand

I establish three descriptive patterns that relate the scale of agricultural production with the aggregate demand for workers in rural economies. The first pattern links this scale with the demand for workers in the agricultural sector, and the other two connect farm size with the demand for nonfarm labor through local consumption. These patterns provide insights into the main channels behind the impact of consolidation on rural employment and motivate the structure of the model that I develop in Section 2.4.

Pattern 1: Agricultural labor intensity declines with farm size

The decline in labor-to-land ratio across farm sizes is one established observation in the development economics literature (Sen, 1981; Carter, 1984; Deininger et al., 2016). This relationship has been explained

[11] Although many retail goods are imported from larger cities, they are almost exclusively consumed within the local economies.

[12] This is perhaps not surprising given that processing plants are more intensive in capital than labor. More importantly, many of the plants that process the main crops such as coffee and oil palm are located in larger cities.

by multiple factors, including access to credit markets and the advantage of smallholders in labor costs (Eswaran and Kotwal, 1986; Sial and Carter, 1996; Foster and Rosenzweig, 2022). Figure 2.3 illustrates this decline in labor intensity using data from the universe of farms in Colombia. This relationship prevails across the country's main seasonal and perennial crops and is particularly salient for crops that are easier to mechanize, such as sugar cane and rice (see Figure 2B in the appendix).[13] Relatedly, data for Colombia suggests that the possession of farm equipment increases sharply with farm size (see Panel B in Figure 2.3). For instance, farmers with more than five hundred hectares are six times more likely to have equipment than farmers with less than three hectares. Thus, to the extent that possession and usage are likely correlated, these findings suggest that the average scale of agricultural production will affect the aggregate demand for workers on the farms.

Pattern 2: Scale economies in non-labor input markets

Figure 2.3 (Panel C) presents information on the costs of farm equipment per hectare for four crops in different regions of the country. In particular, it displays the ratio of the costs per hectare between farms above and below the regional size threshold defining a small farm (i.e., family farm unit –UAF). These costs are associated with the use of motorized equipment to conduct tasks across all production stages. Some of the stages with a higher prevalence of mechanization are harvest and land preparation.[14] The main pattern that emerges from this figure is that large holders spend less money per hectare on farm equipment than smallholders. For instance, large rice producers in the department of Tolima spend 65% of the costs paid by producers with small producers farms. These differences in machinery costs suggest that large producers face a lower rental price in these markets since it is likely that equipment intensity increases with plot size.[15] One important case in point that illustrates these nonlinearities in pricing is tractor services. According to interviews in the field, the rental price per hectare is U$30 for farms with less than one hundred hectares and U$25 (or less) for larger farms.[16] Further, pecuniary economies are also prevalent in other factor markets, such as those for fertilizers and pesticides (DANE, 2022). These observations imply the existence of scale economies in non-labor markets and suggest that aggregate income and consumption depend on the size distribution of farms.

Pattern 3: Non-homothetic consumption growth

Regarding consumption, the expenditure shares in goods and services inside the rural economies have

[13] In addition to the use of motorized equipment to prepare the land, these crops can be harvested using machinery such as the sugar cane cutter and the combine paddy harvester.

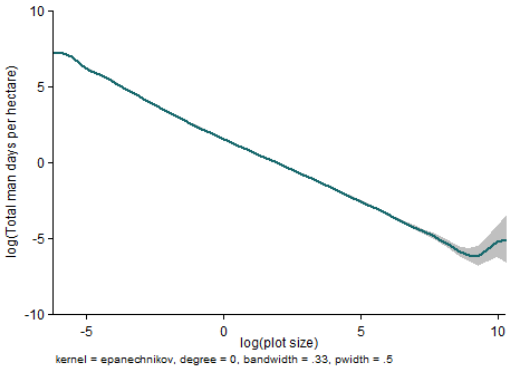
[14] See Perfetti et al. (2012) for details on the relative importance of inputs across the stages of production for these crops.

[15] Formally, let $\frac{p_j k_j}{h_j}$ be the the total cost of equipment per hectare for a farm with size j, where $j =$ small, large. If $\frac{k_s}{h_s} < \frac{k_l}{h_l}$, then $\frac{p_s k_s}{h_s} > \frac{p_s k_s}{h_s}$ implies that $\frac{p_s}{p_l} > \frac{k_l h_s}{k_s h_l} > 1$.

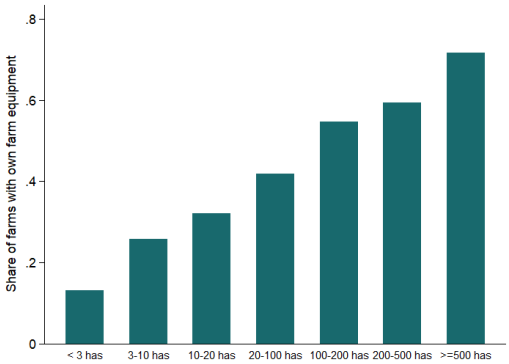
[16] These values were obtained from focus groups during fieldwork in 2019, and are calculated using an exchange rate of $4,000 Colombian pesos.

Figure 2.3: Empirical Patterns Linking Farm Size and Labor Demand

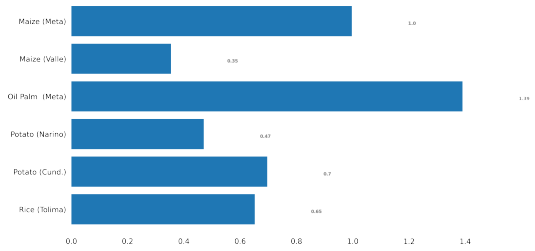
A. Labor Intensity



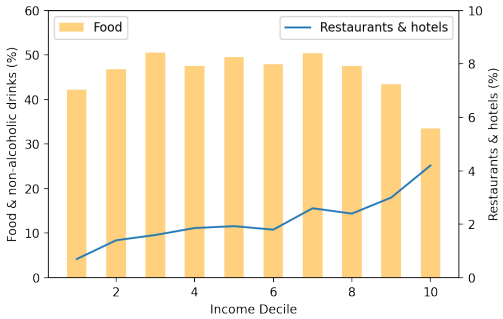
B. Possession of Farm Equipment



C. Machinery Costs per Hectare (Ratio)



D. Expenditure Shares



Notes: Panels A (Panel B) displays the relationship between farm size and labor intensity (farm equipment possession) using data from the agricultural census. Panel C presents the ratio in machinery costs per hectare between small and larger farms using data from the cost analysis conducted by Perfetti et al. (2012). Panel D displays the expenditure shares in i) food and non-alcoholic beverages, and ii) restaurants and hotels by decile of income using data from the national household budget survey. See Section 2A for details on the data and definitions.

been shown to vary across income levels. For instance, using data from rural villages in Mexico, Taylor and Dyer (2009) document a decline in the proportion of expenses on local goods (food + nonfarm) as income per capita in the household increases. Households with an average income of U$356 spend 34% of their budget outside the villages, while this share ascends to 43% for households with double that income. In the absence of granular data on purchases made inside and outside the local economies in Colombia, Figure 2.3 documents a similar pattern of non-homothetic consumption using data on expenditures by industry (see Panel D). On the one hand, the expenditure share in industries related to outside consumption, such as restaurants and hotels, increases with the income decile. Conversely, in line with Engel's law, the expenditure on food displays a downward trend across the right tail of the income distribution. These two observations suggest that the budget share allocated to local nonfarm goods varies across levels of income, and thus, to the extent that large producers are more likely to be wealthier, the demand for labor in the rural nonfarm sector will depend on how concentrated agricultural profits are.[17] This variation in local consumption is potentially exacerbated by the prevalence of absentee landlords in rural regions of the country, which are usually large producers that live most of the year in the capital cities (Adams, 1966; Edel, 1971; Robineau et al., 2010; Piniero, 2016). For instance, Robineau et al. (2010) find that 40% of the farmland in two municipalities of the highlands belongs to absentee owners who hire a manager to take care of production.

2.4 Two-Sector Model of a Rural Economy

I develop a static model of a rural economy with two sectors. This model features a small-open economy where farmers are heterogeneous in their landholdings, input demands vary with plot size, and local consumption is a major determinant of aggregate labor demand. First, I present the economy's endowments, the market environment, and the technologies. Then, I define the competitive equilibrium and explain the intuition behind the comparative statics. The last subsection discusses the role of land sales markets and examines the type of transactions that would take place in the long run.

2.4.1 Environment

Consider a small open economy that produces an agricultural crop q_a and a nonagricultural good q_n. The prices of these two goods are p_a and p_n, respectively, and the non-agricultural good is nontradable. In this economy, individuals consume these two locally produced goods, in addition to a third good that is produced and sold in the city q_c. This city good is a higher-quality substitute for the non-agricultural good that is

[17] For instance, it is reasonable to think that wealthier producers are more likely to replace local goods with higher-quality counterparts produced and sold in the cities (De Janvry and Sadoulet, 1993).

produced locally.

This local economy is endowed with fixed supply of labor L and farmland H. All individuals are endowed with one unit of labor that is supplied inelastically but have different endowments of land. In particular, there are θL agents that have landholdings, and each one of them has a different endowment of hectares given by the distribution $f(h)$. The supply of capital in this rural economy K comes from the city and is perfectly elastic at a fixed rate r. Once landholdings are realized, individuals make their choices in two stages. First, they choose inputs to maximize profits, and then they finance consumption with their disposable income.

In this model, agents are competitive in output markets, and labor is perfectly mobile across sectors. This latter assumption implies a unique equilibrium wage w and precludes selection as a potential channel for the allocation of employment across sectors. Consistent with the patterns documented in Section 2.3, there are pecuniary economies in the agricultural capital market. Namely, the effective rental price that farmers pay per unit of capital R_i varies across agents and decreases with holding size, i.e, $R_i'(h_i) < 0$. The land market in this model also departs from its competitive benchmark. Notably, land cannot be rented in or rented out. This assumption is consistent with the finding that rental markets in Colombia are thin due to high informality and credit rationing (Gáfaro et al., 2014). Thus, this static model is intended to explain how exogenous shifts in the farm size distribution affect rural employment rather than how this new distribution emerges.[18]

Technologies:

Agriculture: the agricultural good is produced using land h, labor ℓ_a, and capital k_a. The technology features constant returns to scale and is given by

$$q_a = h^{\gamma}[\eta \ell_a^{\rho} + (1-\eta)k_a^{\rho}]^{\frac{1-\gamma}{\rho}}$$

where γ is the output elasticity of land and $\eta \in (0,1)$ captures the relative importance of labor to capital in production. The parameter $\rho < 1$ determines the elasticity of substitution between labor and capital, i.e. $\sigma = \frac{1}{1-\rho}$. As $\rho \to 1$, labor and capital become highly substitutable.

Nonagriculture: the nonagricultural good is produced by a stand-in firm. The technology uses labor and capital and features constant returns to scale:

$$Q_n = L_n^{\alpha} K_n^{1-\alpha}$$

[18] See Section 2.4.4 for a discussion on the role of sale markets in the consolidation of land in the long run. For a two-sector static model that endogenizes the farm size distribution, see Adamopoulos and Restuccia (2014) and Chen (2017).

The nonfarm good is nontradable and thereby production depends on local demand. This characterization precludes the production of value-added products to focus attention on the role of local consumption as a driver of employment in this sector (see Section 2.3).[19] In contrast to agricultural production, there are no pecuniary economies in the production of the nonfarm good. Thus, the effective price of capital for the stand-in firm equals the exogenous rate determined in the city, r, and $\pi_n^* = 0$.

Preferences:

I consider an economy where agents consume three types of goods. Two of them are the agricultural $\{c_a\}$ and nonagricultural $\{c_n\}$ goods produced in the rural economy. The third one is a good that is produced and sold in the city $\{c_c\}$, which is a higher-quality substitute for the locally produced (and traded) nonfarm good. Agents choose consumption to maximize

$$\begin{aligned} \max_{c_a, c_n, c_c} \quad & u(c_a, c_n, c_c) = \omega_a log(c_a - \bar{c_a}) + \omega_n log(c_n) + \omega_c log(c_c + \bar{c_c}) \\ \text{s.t.} \quad & p_a c_a + p_n c_n + p_c c_c \leq Y \end{aligned} \tag{2.4.1}$$

where ω_j is the relative weight for good j, $\bar{c_a} > 0$ is a subsistence constraint of food consumption, and $\bar{c_c} > 0$ determines the level of utility when consumption of the city good is zero. The motivation behind these non-homothetic preferences is to account for the observation that changes in income lead to changes in expenditure shares (see the fourth pattern documented in Section 2.3). If $\bar{c_a} > 0$, the income elasticity of the agricultural good is less than one. If $\bar{c_c} > 0$, the income elasticity of the city good is greater than one.[20]

The Agent's Problem:

The problem for each agent can be solved in a recursive way. First, she maximizes income. Then, she chooses consumption demands subject to that income. The disposable income of a landless agent is given by the value of their labor endowment $Y = w$. The disposable income of a farmer is given by the sum of the value of their labor and the agricultural profits, $Y = w + \pi_a^*$.

Given the market environment, agricultural profits are given by:

$$\max_{\ell_a, k_a} \quad \pi_a = p_a q_a - w\ell_a - R_i(h)k_a \tag{2.4.2}$$

[19] According to the National Household Survey, manufacturing of food, beverages, tobacco, and leather products only contributed with 8% of rural nonfarm employment in 2017. This represents a two percentage point increase since 2006.

[20] Note that these preferences are defined as long as $Y \geq p_a \bar{c_a}$

To gain insight, I posit that the effective rental price for each farmer i is

$$R_i(h) = r\Big(1 + \frac{1}{h_i^{\nu}}\Big)$$

where r is the benchmark price that is exogenously determined in the city and $0 < \nu < 1$.[21]

2.4.2 Equilibrium

I focus on the competitive equilibrium of the model. Given the market environment, two prices are endogenously determined. These prices are the equilibrium wage w and the price of the nontradable good p_n. The price of the agricultural crop p_a is set in the international market, and the prices of the city good p_c and capital r are determined in their respective markets in the city.

Market of nonfarm good: the problem for the stand-in firm that produces the non-farm good implies that aggregate production is perfectly elastic. This indicates that total production is pinned down by the aggregate consumption of the good,

$$Q_n = C_n = (1-\theta)Lc_n^0(w) + \theta L\left[\int^{h^{-1}(\widetilde{Y})} c_n^1(h)dF(h) + \int_{h^{-1}(\widetilde{Y})} c_n^2(h)dF(h)\right] \tag{2.4.3}$$

where $c_n^0(w)$ is the consumption function of landless agents, and $c_n^2(h)$ ($c_n^1(h)$) is the consumption for agents who do (do not) purchase the city good.[22] Since $\pi_n^* = 0$, in equilibrium, p_n^* equals the marginal cost $MC(w, r)$.

Labor market: the clearing condition that defines the equilibrium wage in this economy is given by the equality of total labor supply and total labor demand,

$$\bar{L} = L_n + L_a = L_n + \theta L \int \ell_a(h)dF(h) \tag{2.4.4}$$

The aggregate labor demand equals the sum of the demand for workers in the non-farm sector L_n and the total workers required by the farmers L_a. Since the supply of the nonfarm good is perfectly elastic, L_n is pinned down by local consumption C_n.

[21] This functional form implies that $R_i'(h_i) < 0$ and $R_i''(h_i) > 0$.

[22] $\widetilde{Y}$ is the income threshold at which agents start consuming the city good.

Definition 2.4.1. *The equilibrium in this economy is given by input demands* $\{L_a, L_n, K_a, K_n\}$*, consumption demands* $\{C_a, C_n, C_c\}$*, and prices* $\{w, r, p_a, p_n, p_c\}$ *such that:*

1. Input demands in agriculture satisfy the agent's optimization in Equation 2.4.2.
2. Consumption demands are consistent with the agent's optimization in Equation 2.4.1.
3. Total production of the nonfarm good is given by Equation 2.4.3, and $p_n^* = MC(w, r)$.
4. Labor market clears following the equality in Equation 2.4.4.
5. Total farmland equals the sum of endowments across landed agents: $H = \theta L \int h_i dF(h)$.

The exogenous price p_a^* is set internationally and p_c^* and r^* are set in the city.

2.4.3 Characterization and Discussion

Partial Equilibrium Relationships

In what follows, I provide three propositions that summarize the main partial equilibrium relationships of the model. These relationships provide insights into the key ingredients of this framework and are consistent with the empirical patterns presented in Section 2.3. The proofs of these propositions are provided in Section 2C.2.

Proposition 1. *If* $\nu > 0$ *and* $\frac{1}{\sigma} < \gamma$*, labor per hectare decreases with holding size.*

This proposition indicates that if the elasticity of substitution between labor and capital is high, large producers will substitute labor for capital and become less labor intensive than smaller producers. This substitution is driven by the nonlinearities in the pricing of capital for farmers. If $R_i = r$ for all producers, input intensity would not vary across farm size.

Proposition 2. *If* $0 < \nu < 1$*, yields and profit per hectare increase with holding size.*

Proposition 3. *There exists an income threshold* $\widetilde{Y}$ *at which agents start consuming the city good:*

$$\widetilde{Y} = p_a \bar{c_a} + \frac{p_c \bar{c_c}(1 - \omega_c)}{\omega_c}$$

Further, if $p_a \bar{c_a} < p_c \bar{c_c}$ *, there is a hump-shaped relationship between the expenditure share of the local nonfarm good and income.*[23]

[23] In contrast to dynamic models of structural transformation, this static model does not aim to be consistent with a balanced growth path. Therefore, this potential hump-shaped relationship between expenditure share and income is given by the possibility that $p_a \bar{c_a} - p_c \bar{c_c} \neq 0$. Balance growth, on the other hand, requires that $p_a \bar{c_a} - p_c \bar{c_c} = 0$. In this case, the expenditure share of c_n would remain constant as income increases while the expenditure share of c_a (c_n) would decline (increase). See Herrendorf et al. (2014) for more details on the dynamic models that are consistent with balanced growth path.

This proposition indicates that there are two potential consumption regimes in the economy.[24] First, given subsistence level $\bar{c_a} > 0$ and $\bar{c_c} > 0$, individuals with low income levels (i.e. $Y \leq \widetilde{Y}$) set $c_c^* = 0$ and spend their money on food and the locally traded nonfarm good.[25] In this regime, the subsistence constraint ceases to bind as income increases, and individuals start substituting expenditure on food with expenditure on nonfarm goods. Thus, an increase in income leads to an increase in the budget share allocated to the locally produced good, i.e., c_n.

At higher levels of income, $Y > \widetilde{Y}$, individuals can afford to purchase the good produced and sold in the city ($c_c^* > 0$). In this regime, non-homothetic preferences imply that individuals increase their budget share spent on this good as income augments. If $p_a\bar{c_a} < p_c\bar{c_c}$, this budget share increases at a higher rate than the rate at which the share in food expenditures decreases. As a result, when income increases, individuals finance their consumption of the city good by reducing the expenditure share in both food and the nonfarm good.

Mechanisms

In this subsection, I describe the main mechanisms behind the effect of an exogenous increase in large-scale concentration on sectoral employment and the equilibrium wage. To illustrate the intuition, consider Panel A of Figure 2.4, which features the rural labor market. In this figure, the x-axis represents the total supply of workers in the economy, and the red line represents the demand for farm and non-farm labor. The blue line depicts the demand for non-farm labor, which has its origin in the bottom right corner of the figure.

Effect on the equilibrium wage: an increase in large-scale consolidation will have two main effects on the equilibrium wage. First, an increase in the average plot size in the economy will result in a decrease of labor intensity, and thereby a reduction in farm labor demand (i.e. labor intensity effect).[26] Second, consolidation will affect the demand for nonfarm labor since the expenditure in local consumption, $\mathcal{C}_n$, varies with the changes in agricultural profits. In contrast to the labor intensity effect, this shift in the demand for nonfarm labor is ambiguous as it depends on two offsetting effects. On one hand, a larger average plot size increases agricultural profits and thereby the income to spend in the local economy (i.e. profit effect). On the other, the concentration of profits will decrease the income share allocated to local consumption, as wealthier farmers are more likely to replace local consumption with goods that are only sold in the city

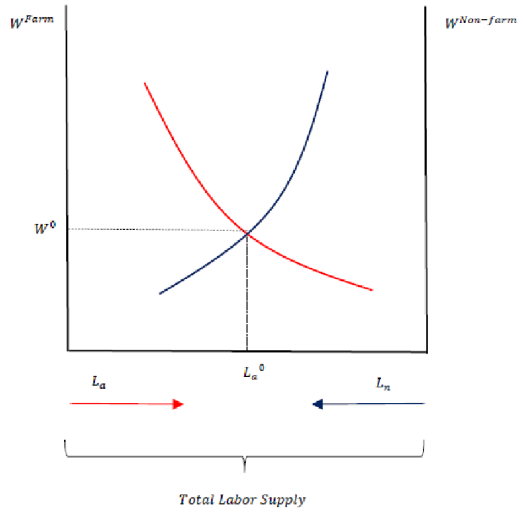
[24] If $\widetilde{Y} <= w$, there is only one regime as everyone can afford to buy the city good.

[25] This corner solution in the consumption of city goods is relevant for rural areas of developing countries due to the vast heterogeneity in individual income.

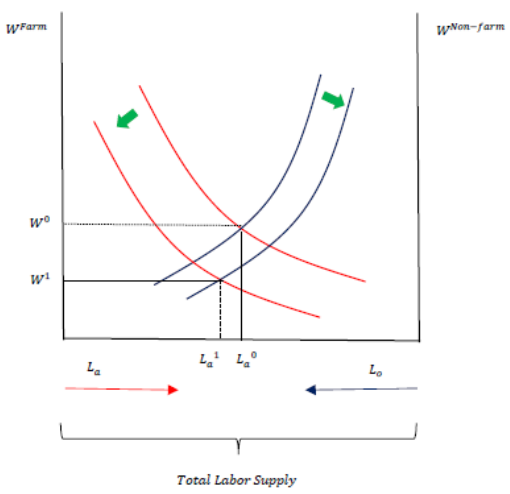
[26] In theory, the direction of the labor intensity effect depends on the elasticity of substitution between labor and capital (see Proposition 1). However, in this illustration, I focus on the case where labor intensity decreases with plot size since this is consistent with the empirical patterns documented for the main crops in Colombia in Section 2.3.

Figure 2.4: Overall Effect on Equilibrium Wage: An Illustration

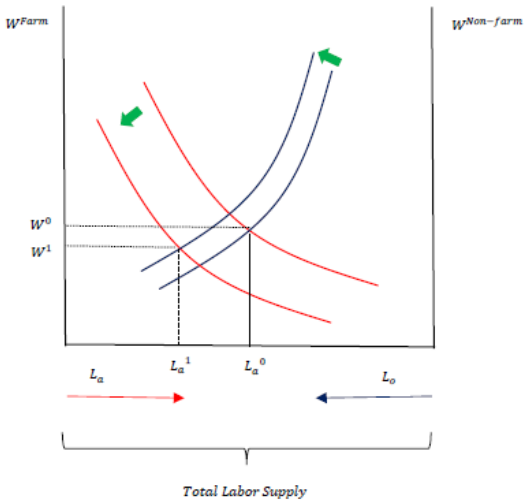
Panel A. Rural Labor Market



Panel B: Negative Pull Effect



Panel C: Positive Pull Effect



Notes: Own illustration.

(i.e. budget share effect). Given these offsetting effects, the shift in the demand for nonfarm labor as a consequence of large-scale consolidation will depend on whether the positive profit effect is larger than the negative budget share effect. If the positive effect is larger, the demand for nonfarm labor will increase (i.e. positive pull response). Conversely, if the budget share effect is larger, there will be a decrease in the nonfarm labor demand (i.e. negative pull response).

To illustrate how these three effects impact the equilibrium wage, refer to Figure 2.4 . If there is a negative pull response in the nonfarm sector, the demand for nonfarm labor will decrease and there will be an *unambiguous* reduction in the equilibrium wage (Panel B). In contrast, if there is a positive pull response, the net effect on the wage will be *ambiguous* and will depend on whether the increase in the demand for nonfarm labor is large enough to offset the labor intensity effect (Panel C).

Effect on the share of agricultural workers: similar to the analysis on the comparative statics for the wage, the effect of an increase in concentration on the share of agricultural workers will depend on the sign and magnitude of the shift in the demand for nonfarm labor. A positive pull response will result in a decrease in the share of farm workers, as the gap in sectoral wages will promote the reallocation of labor from the farm to the nonfarm sector until a new equilibrium emerges. Meanwhile, a negative pull response will result on an ambiguous overall effect since the type of reallocation will depend on the relative strength of the inward shifts in both curves.

Two important insights emerge from this analysis. First, land consolidation affects the demand for both farm and non-farm labor. Second, the sign of the overall effect on the equilibrium wage is ambiguous. This overall effect depends on the interaction of three main competing effects. The labor intensity effect, the profit effect, and the budget share effect.

The last subsection discusses the role of land sales markets and examines the type of transactions that would take place in the long run.

2.4.4 Does Land Get Consolidated Via Sales Markets In The Long Run?

In this model, the distribution of land only changes due to exogenous policies since land cannot be rented in our out. However, in the long run, the distribution of farm size can change through the operation of the sales market. In this section, I analyze the role of these markets following the approach in Carter and Salgado (1998) and Carter and Zegarra (1994). In particular, I compare the competitiveness of landholders using a land valuation exercise. The question at hand is: what type of transactions would take place if competitive sales markets are considered?

Given the model in Section 2.4.1, an agent's willingness to rent land is given by her shadow rental price:

$$\mu_i = \gamma h_i^{(\gamma-1)}[\eta(\ell_i^*)^\rho + (1-\eta)(k_i^*)^\rho]^{\frac{(1-\gamma)}{\rho}}$$

where h_i is the size endowment and ℓ_i^* and k_i^* are the optimal demands for labor and capital, respectively.[27] Thus, we can obtain a measure of land valuation for each agent i as the capitalized stream of the income increments given by this shadow rental price:

$$\Delta_i = \sum_{t=1}^{T} \frac{\mu_{it}}{[1+r]^t}$$

where r is the interest rate and T is the time horizon over which the household expect to receive benefits from that land. This measure represents the net present production value of land for each agent. Notably, if a household wants to buy (sell) one unit of land, Δ_i is the maximum (minimum) price that agent i is willing to pay (accept) for that unit without losing money. The higher this value, the more competitive an agent is in the sales market as land will flow to those agents with a higher willingness to pay.

Proposition 4. *If $\mu_{it} = \mu_i$ and $\gamma < 1$, the net present production value of land increases with plot size*

$$sign\Big(\frac{\partial \Delta_i}{\partial h}\Big) = sign\Big(\frac{\partial \mu_i}{\partial h}\Big) > 0$$

The proposition above implies that large producers place a higher value on land. In particular, given advantages in capital markets, they are more productive and thereby more willing to pay a higher price than small producers for one unit of land. This difference in the valuation of land makes large holders more competitive in sales markets. Thus, since the land will flow from smaller to larger producers, these results suggest that the model would predict a consolidation in large farms if sales markets are considered.

This net present value approach to land valuation is useful and informative about the core income factors that shape the willingness to pay for land. Yet, in comparison to a dynamic general equilibrium model, it abstracts away from risk and intertemporal consumption choices that may also affect the willingness to pay for land (Carter and Kalfayan, 1989; Carter and Salgado, 1998; Carter and Zegarra, 1994). While these two factors may have opposing effects on the competitiveness of smallholders, previous research shows that their consideration may reduce their land valuation even further (Carter and Zimmerman, 2000). Thus, the results obtained with a dynamic model will probably enforce the long-run tendency to consolidate land in large farms.

[27] Note that if rental markets were thick and considered in the modeled, an agent's willingness to pay to rent land would be equal to the market rental price.

2.5 Model Calibration

2.5.1 Parameters and Targets

I calibrate the model to data on Colombia's rural economy. The parameters to calibrate are: technological parameters $(A, \gamma, \rho, \eta, \nu, \alpha)$, preference parameters $(\omega_a, \omega_n, \bar{c_a}, p_c\bar{c_c})$, distributional parameters (θ, τ), exogenous prices (p_a, r), and endowments (L, H). The calibration strategy follows two steps. First, I calibrate some parameters externally based on values taken from the literature. Some of them are assigned directly, and others are chosen to match analytical properties of the farm size distribution that are independent of the model's equilibrium outcomes. The second step consists of using the structure of the model to jointly calibrate the remaining parameter values by matching targeted moments from the model to those in the data.

Parameters calibrated externally: Table 2.1 summarizes the parameters that are calibrated based on values from the literature. I normalize productivity and the output price in agriculture (A, p_a) to 1 and set $\gamma = 0.40$ based on the land shares estimated by Avila and Evenson (2010).[28] Similarly, I set the labor share in nonfarm production to $\alpha = 0.59$ based on the estimates for services in Hamann et al. (2019). The relative weights of consumption goods in preferences are chosen following the structural transformation literature. On one hand, I set ω_a equal to the long-run share of rural employment in agriculture.[29] On the other, I set ω_n based on the estimates of the weight of services in preferences obtained by Herrendorf et al. (2013).[30] The motivation for using estimates from the service sector to calibrate ω_n is that the local nonagricultural good in the model is non-tradable.

In regards to land, I set the proportion of individuals with land to $\theta = 0.42$ based on the estimates of Gáfaro et al. (2014). Further, since the distribution of farm sizes resembles a Pareto distribution, I assume that $F(h) = 1 - \left(\frac{h_m}{h}\right)^{\tau}$, where τ refers to the shape parameter and h_m corresponds to the minimum farm size.[31] To create the grid of farm sizes, I use a discrete approximation of the Pareto distribution following the approach in (Henderson and Isaac, 2017). I set $\tau = 1.08$ and $h_m = 1.29$ to match a Gini of 0.86 and an average farm size of 17.2 hectares.[32] Additionally, given the total number of farmland hectares ($H = 31$

[28] This value is obtained as the area-weighted average of the land share in crop production (0.23) and livestock raising (0.44) in Colombia. The value of land share for crops is in line with that of Peru (Sotelo, 2020), which has a similar crop portfolio.

[29] This share is obtained using data from the 1990 US population census since it is the latest source that reports these statistics for rural areas. This value (7%) is larger than the national long-run share of agricultural workers used in the literature, which oscillates between 1% and 5% (Chen, 2017; Adamopoulos and Restuccia, 2014; Herrendorf et al., 2013).

[30] The results of Herrendorf et al. (2013) imply that services account for 83% of the importance of the nonagricultural goods in preferences. Thus, since $\omega_a = 0.07$, I calculate ω_n as $0.83 \times (1 - 0.07) = 0.77$.

[31] The Pareto distribution is a standard way to model land heterogeneity in the development literature. See, for instance, Bazzi (2017), Carter and Kalfayan (1989), and Eswaran and Kotwal (1986).

[32] This average farm size value is calculated using data from the agricultural census and is similar to the one estimated by Hamann et al. (2019).

million), I set the rural workforce L to match a farm-to-labor ratio of 7.22. This ratio is obtained from a property of the Pareto distribution, $\frac{H}{L} = \theta \frac{\tau h_m}{(\tau - 1)}$, and yields a workforce value that is close to recent estimates from national household surveys (Otero-Cortés, 2019).

Finally, I use data on the real interest rate to determine the value of the benchmark price of capital. Specifically, I set $r = 7\%$ based on the average rate over the last decade.

Table 2.1: Parameters Calibrated Externally

Description	Paramater	Value
Productivity in agriculture	A	1.000
Price agricultural good	p_a	1.000
Land share of income in agriculture	γ	0.398
Labor share of income in nonfarm sector	α	0.590
Weight agricultural consumption	ω_a	0.072
Weight local nonfarm consumption	ω_n	0.770
Share of landed agents	θ	0.416
Pareto distribution (shape)	τ	1.081
Pareto distribution (scale)	h_m	1.294
Aggregate farmland	H	31,046
Aggregate labor	L	4,300

Notes: This table presents the parameter values taken from the literature or obtained from properties of the Pareto distribution. Aggregate land refers to thousands of hectares, and aggregate labor corresponds to thousands of rural workers.

Parameters calibrated using the structure of the model: Equipped with the distribution of farm sizes and the parameters in Table 1, I jointly choose the value of the remaining five parameters $(\eta, \rho, \nu, \bar{c_a}, \bar{c_c})$ to match five moments in the data.[33] These moments are the labor cost share in agricultural production, the ratio of labor per hectare between smallest and largest farm, the average ratio of capital's price between large and small farms, the expenditure share in agricultural goods, and the expenditure share in nonagricultural tradables.

Although these five parameters affect all moments simultaneously, some moments are more informative of certain parameters. On the one hand, farm costs and input intensity are more important to identify the parameters pertaining agricultural production since the latter determine input choices. For instance, the target on labor intensity ratio across farm sizes is informative of the elasticity of substitution between labor and capital since the higher the ratio the higher ρ is. Additionally, the target on the capital's price ratio across farm sizes is informative of ν since in the model the capital's price ratio between any two farm sizes

[33]This calibration is conducted using the simulated method of moments procedure, which obtains parameter vector b by minimizing the distance between the model simulated moments $m(b)$ and the corresponding set of actual moments in the data.

i and j is:[34]

$$\frac{R_i}{R_j} = \frac{\left(1 + \frac{1}{t_i^{\nu}}\right)}{\left(1 + \frac{1}{t_j^{\nu}}\right)}$$

On the other hand, expenditures in food and tradables are more important to identify preference parameters since $\bar{c_a}$ and $\bar{c_c}$ affect consumption levels. A key feature of the model is that expenditure in consumption is proportional to the quantity consumed as all consumers face the same price. Thus, in the absence of price data on nonfarm city goods (p_c), I use the following transformation of the preferences and jointly identify $p_c\bar{c_c}$.[35]

$$u(c) = \omega_a log(c_a - \bar{c_a}) + \omega_n log(c_n) + \omega_c log(p_c c_c + p_c \bar{c_c}) - log(p_c)$$

The value of the targeted moments are either taken from the literature or calculated using surveys that are representative from the rural context in Colombia. The labor intensity ratio is calculated using data from the 2014 National Agricultural Census, and the labor cost share is obtained from the estimates in Avila and Evenson (2010) by averaging the values across crop and livestock farming. In regards to expenditures, I obtain measures on the share of final expenses using aggregate data on monthly household spending from the Households' Budget Survey. The share in agricultural goods refers to the expenditures in food and non-alcoholic drinks, while the share in nonfarm tradables refers to expenditures in home furniture, clothing, footwear, alcoholic drinks, and restaurants and hotels.[36] Finally, the capital price ratio is calculated as the average price for tractor services between farms above and below one hundred hectares (U$25 and U$30 per hectare, respectively). These values were obtained from focus groups conducted during fieldwork in the second semester of 2019 since there are no official surveys that collect this data in the country.[37]

Table 2.2 presents the value of the parameters obtained with the calibration as well as the value of the target moments in the model and the data. The target moments in the data are matched closely by the corresponding moments in the model. The results suggest that $\eta = 0.89$, which is close to the value calculated by Foster and Rosenzweig (2022) in their analysis of productivity and farm scale in India. In addition, the findings imply a high substitution of labor and capital in farm production ($\rho = 0.99$), which is consistent with the empirical pattern of decreasing labor intensity across farm sizes. Given the distribution of farm

[34] In practice, one alternative moment that is informative of the decay in capital pricing ν is the capital cost ratio between large and small farms. As discussed in Section 2.3, this ratio has an average value of 0.86 suggesting that there are pecuniary economies in capital markets. In the model, however, the calibrated elasticity of substitution between labor and capital is high and capital per hectare increases with farm size. Thus, since large farmers use much more capital than small producers, it is not feasible to match the target value by design.

[35] See Section 2C.3 in the appendix for more details on this transformation of the preferences.

[36] Note that these measures are obtained using the final consumption framework, which is more appropriate when the dominant force in the model is the income effect (and not the price effect) Herrendorf et al. (2013).

[37] The values are calculated using an exchange rate of $\$4,000$ colombian pesos. The study on production costs by Perfetti et al. (2012) discussed in Section 2.3 does not collect price data.

sizes, the value $\nu = 0.07$ in the model suggests that the effective capital price paid by the largest farm is two-thirds of the price paid by the smallest farm. The level of subsistence consumption $\bar{c_a}$ that matches expenditure shares in agricultural goods is 0.72, implying that the income elasticity of the agricultural good is smaller than one.[38] On the other hand, since $p_c\bar{c_c} = 0.00003 > 0$, the income elasticity of the city nonfarm good is greater than one.

This set of calibrated parameters yields a proportion of rural agricultural workers that is lower than the observed one in the data, however. Specifically, the proportion in the model is 0.30 while the data suggests is 0.47 (Tenjo and Jaimes, 2015).[39] To reconcile the proportion in the model with that in the data, I introduce a barrier to labor mobility between the agricultural and nonagricultural sectors in rural economies. In particular, I assume that the standing firm that produces the nonfarm good needs to pay extra benefits for its employees (e.g. contributions to health and retirement).[40] These benefits are a portion $0 < \psi \leq 1$ of the wage, so the firm's unitary cost of labor is $w(1+\psi)$. I assume they are paid to the national government, and thus they are not spent in the local economy. This barrier is chosen to match a proportion of agricultural workers of 0.47 and remains unchanged in the quantitative analyses in Section 2.6.

Table 2.2: Parameters Calibrated Using The Structure of The Model

Paramater	Value	Moment	Target	Model
η	0.885	Labor cost share	0.36	0.34
ρ	0.989	Labor-to-land ratio (smallest/largest farm)	240	238
ν	0.071	Capital price ratio (large/small farms)	0.83	0.83
$\bar{c_a}$	0.716	Expenditure share in agriculture	0.31	0.30
$p_c\bar{c_c}$	3e-05	Expenditure share in nonagricultural tradables	0.13	0.12
ψ	0.989	Share of agricultural labor	0.47	0.41

Notes: This table presents the parameter values that are obtained with the simulated method of moments procedure. The last two columns report the value of the target moment in the data and the corresponding value in the calibrated model.

2.5.2 Model Performance and Discussion

The calibrated model does a good job matching untargeted features of the land distribution in Colombia. Although the calibration only targets the overall Gini index assuming a Pareto density, the model generates a distribution of farm size that resemble the observed one in the data (see panel A in Figure 2.5). For instance, it reproduces the observations that close to seventy percent of farms are smaller than four hectares.

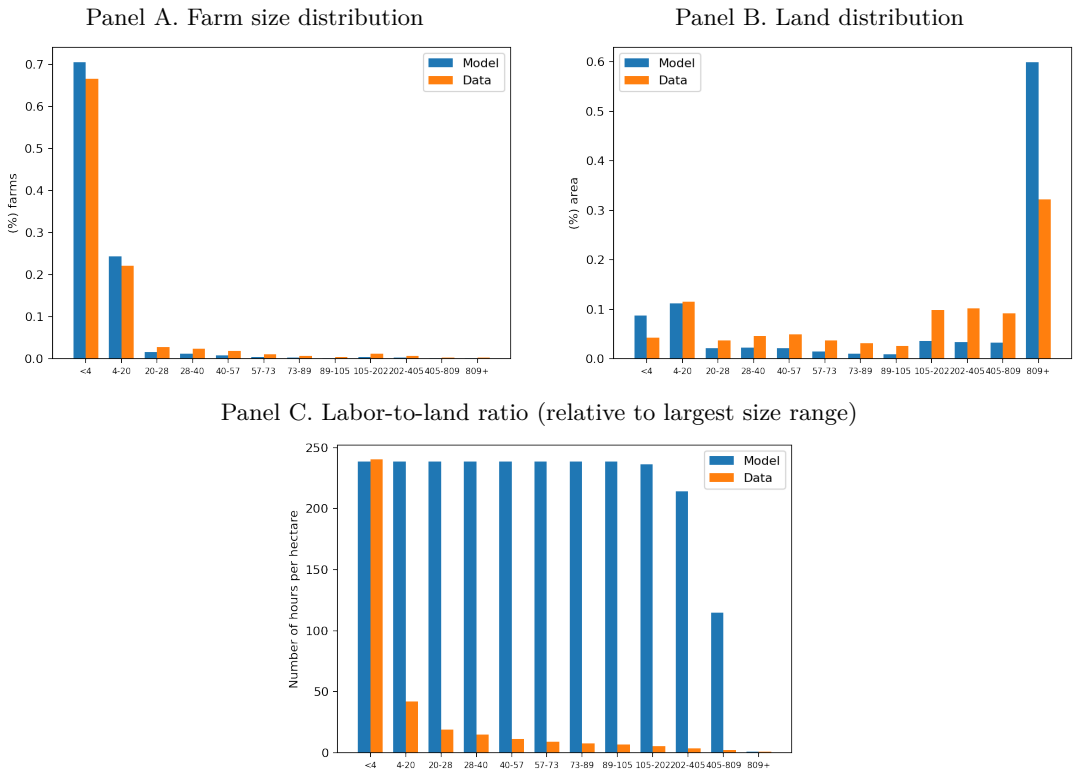
[38] This value of $\bar{c_a}$ is higher than the one obtained in Hamann et al. (2019), which is consistent with the higher consumption of agricultural goods in rural areas.

[39] This proportion is obtained using the new classification of rural towns, which represents a more precise approximation to the rurality of the country. The figures provided in Section 2.2.3 use the standard rural definition since the microdata of the household surveys with the new classification are not publicly available.

[40] This assumption is inline with the observation that the nonagricultural sector has a higher prevalence of formal jobs. For other studies that introduce a barrier to labor mobility across sectors to reconcile moments in the model and the data see Chen (2017) and Adamopoulos and Restuccia (2014).

Similarly, the calibrated distribution of area resembles the shape of the observed one (see panel B in Figure 2.5), indicating that farms over one hundred hectares account for more than sixty percent of farmland. The Pareto density, however, tends to overestimate the area in farms that belong to the last size category. As a result, large-farm consolidation – defined as the share of area in farms above 500 hectares– is somewhat higher in the model relative to the data (61% vs. 40%).[41]

Figure 2.5: Untargeted Variables by Farm Size



Notes: These figures present the value of untargeted variables in the model with their respective counterpart in the data. The values in the data are calculated using the 2014 National Agricultural Census. See Section 2A for more details on the census and definitions.

Regarding input use, the model is qualitatively consistent with the negative relationship between labor intensity and farm size in the data. However, in contrast to the distribution of land, it does not successfully match the labor intensity ratios across *untargeted* size categories. As shown in panel C in Figure 2.5, the model fails to capture the sharp decrease in labor intensity for farms over four hectares and overestimates the labor-to-land ratio for farmers across most size categories. These discrepancies are explained by the high elasticity of substitution between labor and capital in the model since $\rho = 0.99$ implies that both inputs are close to being perfect substitutes in farm production. Note that ρ affects the aggregate demand for agricultural workers besides the labor intensity ratio between the largest and smallest size categories. Thus,

[41] This is perhaps not surprising given that the Pareto density Type I is characterized by having a long tail. In a future version of this work, I will calibrate the model to match the measure of large-farm consolidation instead of the overall Gini index to assess the robustness of the results in Section 2.6.

since this ratio is substantially high at 240 hours per hectare, the model requires that most farmers belong to the most labor-intensive regimes in order to attain the target proportion of farmworkers. A possible interpretation of this finding is that ρ is also capturing other factors that drive the substitution of capital and labor in production, such as access to credit markets. Thus, for a future version of this draft, I will adjust the model to consider the role of credit constraints on input choices (Carter, 1984) and use data on credit access from the agricultural census to calibrate it.

2.6 Policy Analysis

I now use the calibrated model to examine how large-farm consolidation affects the structure of rural employment and workers' income. In addition, I study how these effects differ by type of consolidation. I examine two counterfactual policies that attain the same level of large-farm consolidation by merging land from different parts of the distribution. One consolidates exclusively small farms into large farms, while the other one only consolidates land from middle-size farms.

For each policy, I quantify impacts on structural change and workers' income. In addition, I quantify effects on aggregate social welfare using the following Benthamite welfare function:[42]

$$W = N_0 U(c) + N_1 \int U(c) dF(h)$$

where $N_0 = (1-\theta)L$ corresponds to the individuals without access to land and $N_1 = \theta L$ corresponds to landed individuals.

One important caveat of these analyses is that the current calibrated model underestimates labor intensity decreases relative to those observed in the data. As discussed in Section 2.5.2, this is because of the mismatch in labor-to-land ratios across the middle section of the farm size distribution. This mismatch implies that the calibrated model underestimates the decrease in farm labor demand when land consolidation increases. Thus, while these counterfactual exercises are informative about the potential effects of consolidation policies, the results are still preliminary and should be regarded with caution.

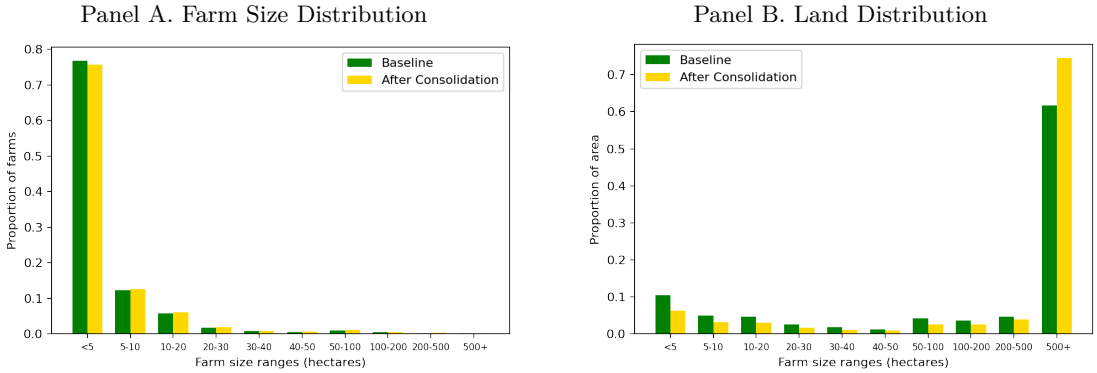
2.6.1 Large-Farm Consolidation and Rural Labor Markets

To quantify the effects of large-farm consolidation, I conduct an experiment that increases the share of area in farms over 500 hectares while holding the economy's endowments and parameters constant. Assuming that the ex-post distribution remains Pareto, I reduce the proportion of landed individuals as to increase

[42] This function has been previously used to study welfare effects in the literature. See, for instance, the analysis in Eswaran and Kotwal (1986).

large-farm consolidation by thirteen percentage points. This increase corresponds to one standard deviation of the change in consolidation observed in the country during the 1990s and constitutes a useful reference to compare the results of the current analysis with the reduced-form estimates from that setting.[43] Figure 2.6 presents the farm size and area distributions before and after the policy. This policy change increases average farm size by eleven hectares and reduces the proportion of landed individuals by sixteen percentage points (see Panel A in Table 2.3).

Figure 2.6: Farm Size Distribution Before and After Large-Farm Consolidation



Notes: These figures present the distribution of farms and land before and after the experiment described in Section 2.6.1. Panel A presents the number of farms across size ranges and Panel B presents the total area in each size range.

The results of this policy experiment indicate that large-farm consolidation has a meaningful impact on the structure of rural employment (see Panels B and C in Table 2.3). The increase in the share of area in large farms led to a four percentage points decline in the proportion of farmworkers. This finding is driven by a decrease in the demand for agricultural labor when the average farm size in the economy increases since the labor to land ratio declined by 10%. In line with the economies of scale in the model, consolidation increases aggregate agricultural profits and the demand for non-tradable services in the economy. As a result, the demand for nonfarm workers at baseline prices increased by 2%.[44] However, since the positive pull effect in demand for nonfarm workers was smaller than the push effect in the farm sector, land consolidation led to a decline in the aggregate demand for rural labor and a reduction in the equilibrium wage of 8%. These results imply that the structural transformation of rural economies was accompanied by a decline in the income of rural workers, which is unexpected given the positive correlation between income and labor reallocation documented in the structural change literature (Herrendorf et al., 2014).

When I focus on the impacts of consolidation on welfare, I find that consolidation led to a slight increase

[43] See Ortiz-Becerra (2022) for more details on the large-farm consolidation that occurred in several regions during the 1990s and its impacts on the income of workers and the structure of rural employment.

[44] This value is obtained by evaluating the demand for nonfarm workers in the ex-post economy at the prices of the baseline economy.

Table 2.3: Impacts of Large-Farm Consolidation

Panel A: Land Distribution

Description	Baseline	Ex-Post	Change
Share of landed (%)	42.0	25.9	-16.0pp
Gini (index)	0.86	0.91	0.05
Average farm size (hectares)	17.2	27.8	10.6
Share of area in large farms (%)	61.0	74.0	13.0pp

Panel B: Labor Markets

Description	Baseline	Ex-Post	Change
Share of farm workers (%)	41.5	37.3	-4.2pp
Equilibrium wage ($)	1.2	1.1	-7.7%
Push vs Pull Effects:			
# farm workers at baseline prices	1,784	1,170	-34.5%
# nonfarm workers at baseline prices	2,516	2,559	1.7%

Panel C: Additional Outcomes

Description	Baseline	Ex-Post	Change
Aggregate farm labor to land ratio	0.06	0.05	-10.0%
Aggregate farm profits ($)	4,149	4,435	6.9%
Production of nontradables (units)	9,182	9,514	3.6%
Consumption (units):			
Aggregate consumption	14,393	14,700	2.1%
Average among landed	4.5	7.2	43.8%
Average among landless	2.2	2.1	-2.7%
Income ($):			
Aggregate income	9,265	9,154	-1.2%
Average among landed	3.5	5.1	45.4%
Average among landless	1.2	1.1	-7.7%

Notes: This table presents the results of the policy experiment described in Section 2.6.1. The second column presents the outcomes of the baseline economy and the third one the economy after consolidation. The change in average farm size refers to number of hectares and the change in Gini refers to the index's units.

in aggregate consumption due to the rise in agricultural productivity and the decline in the price of non-tradables. Yet, a key finding that emerges from this analysis is that these welfare effects vary widely in magnitude and direction across population groups. Landed individuals had a large increase in average income (45%) and consumption (44%), while landless individuals experienced a small decrease in consumption due to the decline in the equilibrium wage. These heterogeneous effects imply that the large majority of rural inhabitants (74%) are worse off after land consolidation, while only a small proportion benefit greatly from it. As a result, aggregate social welfare falls in 53%.

Overall, these findings are qualitatively similar to the quasi-experimental results obtained in Ortiz-Becerra (2022). In that companion paper, I estimate that regions with an increase in large-farm consolidation during the 1990s experienced a ten percentage point decline in the proportion of farmworkers, a fourteen percentage point increase in unemployment, and a general reduction in workers' income of between one and fifty percent. The quantitative model developed in this study explains 42% of the estimated reduction in farm workers despite that the calibration did not target any of these quasi-experimental results. In regards to income, the model explains 15% of the decline in the worst-case estimated scenario and overstates the effects of the most conservative estimates. This potential overestimation is not surprising when we consider that the reduced-form estimates correspond to findings over the span of twelve years while the model is designed to quantify short-term effects in a benchmark economy with no labor mobility across space. Taken together, these results suggest that the model helps explain an important component of the impacts of land consolidation on rural labor markets. Nonetheless, future work is needed to examine the role of migration in the middle and long term.

2.6.2 Does The Type of Consolidation Matter?

I conduct two counterfactual experiments to examine whether the effects of large-farm consolidation vary by the size of the consolidated farms. These experiments are consistent with an exogenous land policy that targets a specific set of farm sizes instead of a market-driven process that affects the whole farm size distribution. In the first experiment, I increase the area in farms with more than five hundred hectares by merging land from the smallest farms into one single large plot. In the second experiment, I increase large-farm consolidation by merging land from middle-size farms into one single large plot.[45] In contrast to Section 2.6.1, I examine an increase in large-farm consolidation of eight percentage points to focus on combining plots that are not too close in size to the large farms.[46] Figures 2C and 2D in the appendix

[45]Holdings between 50 and 500 hundred hectares are defined as middle-sized farms. See Section 2.2.2 for more details on these definitions.

[46]An increase in large-farm consolidation of thirteen percentage points in the second experiment implies the merging of farms that are too close to the five hundred hectares threshold that defines large farms.

present the change in the distribution of sizes and area after each policy.[47]

The first policy has substantial effects on the distribution of land (see Panel A in Table 2.4). It reduces the proportion of landed individuals from 42% to 19% and increases the average farm size by 22 hectares. These significant changes are driven by the fact that many small farms need to be consolidated to achieve a high increase in the proportion of area in large farms. In contrast, the consolidation of middle-sized farms (second policy) has a relatively small effect in these moments of land distribution. The proportion of landed individuals decreases by less than one percentage point, and the average farm size in the economy only increases by 0.23 hectares. This second policy also has a milder impact on overall inequality as measured by the Gini index. Merging middle-sized farms increases the coefficient by 0.01, while combining the smallest farms increases it by 0.02.

Table 2.4 presents the effects of both counterfactual policies relative to the baseline economy. In line with the results in Section 2.6.1, these findings indicate that both consolidation experiments lead to the reallocation of workers out of the agricultural sector and a decline in their income (see Panel B). Interestingly, the magnitudes of these effects are relatively similar across policies. For instance, the wage declines by -2.60% when merging small farms and -2.47% after combining middle-sized farms. These results are unexpected given the stark difference in average farm size across policies and urge the need for further exploration. One possible reason behind this finding might be the limited ability of the calibrated model to reproduce the variation in the labor-to-land ratio across size ranges.[48] In particular, since the model underestimates the changes in labor intensity across the lower tail of the distribution, it is likely that the decline in farm labor demand after merging small farms is undervalued.

With this caveat, the general pattern that emerges from this analysis is that merging smaller farms leads to a slightly larger increase in agricultural productivity and production of non-tradables relative to combining middle-sized farms. This result can be explained by the substantial rise in the average farm size (i.e., scale economies) and the faster increase in the expenditure on local services by farmers with lower income. *Policy 1*, however, is also characterized by a sharper decline in the demand for labor on the farms. Thus, despite the greater gains in productivity, the consolidation of smaller farms leads to a steeper decline in the aggregate demand for rural workers and their wages.

As for welfare, both experiments suggest that the effects on income and consumption are positive for agricultural producers and negative for individuals without access to land. Yet, given differences in the proportion of landed individuals across policies, the average welfare gains for farm producers vary substantially between them. For example, *Policy 1* yields an average consumption increase of 75%, whereas the rise

[47] Note that by definition, land distribution is no longer Pareto after consolidation.
[48] See Section 2.5.2 for a discussion on this limitation.

Table 2.4: Impacts of Large-Farm Consolidation by Merge Type

Panel A: Land Distribution

Description	Base line	Merging Small Farms		Merging Middle-Sized Farms	
		Ex-Post	Change	Ex-Post	Change
	(1)	(2)	(3)	(4)	(5)
Share of landed (%)	42.0	18.5	-23.5pp	41.4	-0.6pp
Gini (index)	0.86	0.88	0.02	0.87	0.01
Average farm size (hectares)	17.2	39.1	21.9	17.4	0.23
Share area in large farms (%)	61.0	69.0	8.0pp	69.0	8.0pp

Panel B: Labor Markets

Description	Base line	Merging Small Farms		Merging Middle-Sized Farms	
		Ex-Post	Change	Ex-Post	Change
	(1)	(2)	(3)	(4)	(5)
Share of farm workers (%)	41.48	40.31	-1.17pp	40.38	-1.10pp
Equilibrium wage ($)	1.19	1.16	-2.60%	1.16	-2.47%
Push vs Pull effects:					
# of farm workers at baseline prices	1,784	1,459	-18.21%	1,475	–17.33%
# of nonfarm workers at baseline prices	2,516	2,522	0.24%	2,522	0.22%

Panel C: Additional Outcomes

Description	Base line	Merging Small Farms		Merging Middle-Sized Farms	
		Ex-Post	Change	Ex-Post	Change
	(1)	(2)	(3)	(4)	(5)
Aggregate farm labor to land ratio	0.057	0.056	-2.81%	0.056	-2.66%
Aggregate farm profits ($)	4,149	4,221	1.73%	4,217	1.64%
Production of nontradables (units)	9,182	9,264	0.90%	9,260	0.85%
Price of nontradables ($)	0.78	0.76	-1.54%	0.77	-1.47%
Consumption (units):					
Aggregate consumption	14,393	14,461	0.47%	14,457	0.45%
Average among landed	4.97	8.72	75.36%	5.07	2.02%
Average among landless	2.17	2.15	-0.91%	2.15	-0.87%
Income ($):					
Aggregate income	9,265	9,203	-0.66%	9,206	-0.63%
Average among landed	3.49	6.47	85.67%	3.53	1.14%
Average among landless	1.19	1.16	-2.60%	1.16	-2.47%

Notes: This table presents the results of the policy experiments described in Section 2.6.2. Column (1) reports the outcomes in the baseline economy. Columns (2) and (3) report the outcomes after the consolidation of small farms (policy experiment 1), and columns (4) and (5) report the outcomes after the consolidation of middle-sized farms (policy experiment 2). The change in average farm size refers to number of hectares and the change in Gini refers to the index's units.

after *Policy 2* is less than 3%. This stark difference in consumption effects across policies has important implications on welfare. Merging the smallest farms decreases aggregate social welfare by 34%, which is almost third times the effect of combining middle-sized farms instead (12%).

2.7 Conclusion

I develop a tractable framework to quantify the effects of large-farm consolidation on the welfare of rural populations. This analysis is particularly relevant in the Colombian setting given the current levels of farmland consolidation and the purpose of the upcoming land reform. Considering the role of i) the non-tradable nonfarm sector and ii) non-homothetic consumption growth enables me to shed light on the distributional effects across workers and producers. This tractable framework also allows me to examine whether these welfare effects vary by type of consolidation.

My findings show that land consolidation affects farmers and workers differently. Farmers benefit from increases in agricultural productivity, while workers are adversely affected by an overall decrease in rural labor demand. Overall, social aggregate welfare declines since large-farm consolidation results in a substantial increase of rural landless. The magnitude of this decline in welfare, however, depends on whether the rise of large operations is driven by merging the smallest farms or combining middle-sized plots.

While this paper is a step toward understanding the welfare effect of large-farm consolidation, it opens several questions for future research. My quantitative framework constitutes a short-term benchmark that abstracts from several factors that are critical when examining land allocation and local labor markets. For instance, spatial labor mobility is an essential margin of adjustment in the medium and long-term (Dix-Carneiro and Kovak, 2019). Moreover, land sales, coupled with imperfections in land markets, can revert the impacts of land policies (Carter and Zegarra, 1994; Carter and Salgado, 1998). Finally, it is crucial to complement this quantitative analysis with (quasi) experimental evidence on the welfare effects of land policies. Such an analysis would allow for the validation of the model and aid in testing the importance of its mechanisms.

2.8 References

T. Adamopoulos and D. Restuccia. The size distribution of farms and international productivity differences. *The American Economic Review*, 104(6):1667–1697, 2014. ISSN 00028282. URL http://www.jstor.org/stable/42920862.

T. Adamopoulos and D. Restuccia. Land reform and productivity: A quantitative analysis with micro data. *American Economic Journal: Macroeconomics*, 12:1–39, 2020. ISSN 1945-7707. doi: 10.1257/mac.20150222.

T. Adamopoulos, L. Brandt, J. Leight, and D. Restuccia. Misallocation, selection and productivity: A quantitative analysis with panel data from china misallocation, selection and productivity: A quantitative analysis with panel data from china †, 2019.

D. Adams. Colombia's Land Tenure System: Antecedents and Problems. *Land Economics*, 42(1):43–52, 1966.

D. Ali, K. Deininger, and A. Harris. Does Large Farm Establishment Create Benefits for Neighboring Smallholders? Evidence from Ethiopia. *Land Economics*, 95(1):71–90, 2019. URL https://ideas.repec.org/a/uwp/landec/v95y2019i1p71-90.html.

J. Arteaga, C. C. Osorio, D. Cuéllar, A. M. Ibañez, R. Londoño Botero, M. Murcia, J. Neva, Á. Nieto, D. I. Rey, and F. Sánchez. Fondo de Tierras del Acuerdo Agrario de la Habana: Estimaciones y Propuestas Alternativas. Documentos CEDE 015630, Universidad de los Andes - CEDE, June 2017. URL https://ideas.repec.org/p/col/000089/015630.html.

A. F. D. Avila and R. E. Evenson. Chapter 72 total factor productivity growth in agriculture. the role of technological capital, 2010. ISSN 15740072.

Á. Balcázar. Transformaciones en la agricultura colombiana entre 1990 y 2002. *Revista de Economía Institucional*, 5(9):128–145, 2003. ISSN 0124-5996.

S. Bazzi. Wealth heterogeneity and the income elasticity of migration. *American Economic Journal: Applied Economics*, 9:219–255, 2017. ISSN 19457790. doi: 10.1257/app.20150548.

A. R. Berry and W. R. Cline. *Agrarian Structure and Productivity in Developing Countries*. Johns Hopkins University Press, Baltimore, 1979.

P. Bustos, G. Garber, and J. Ponticelli. Capital Accumulation and Structural Transformation*. *The Quarterly Journal of Economics*, 135(2):1037–1094, 01 2020. ISSN 0033-5533. doi: 10.1093/qje/qjz044. URL https://doi.org/10.1093/qje/qjz044.

M. Carter. Identification of the Inverse Relationship between Farm Size and Productivity : An Empirical Analysis of Peasant Agricultural Production. *Oxford Economic Papers*, 36(1):131–145, 1984.

M. Carter and J. Kalfayan. A general equilibrium exploration of the agrarian question. Technical report, 1989.

M. R. Carter and R. Salgado. Land market liberalization and the agrarian question in latin america. 1998.

M. R. Carter and E. Zegarra. Land markets and the persistence of rural poverty : Post-liberalization policy options. 8:125–152, 1994.

M. R. Carter and F. J. Zimmerman. The dynamic cost and persistence of asset inequality in an agrarian economy. *Journal of Development Economics*, 63(2):265–302, 2000. ISSN 0304-3878. doi: https://doi.org/10.1016/S0304-3878(00)00117-6. URL https://www.sciencedirect.com/science/article/pii/S0304387800001176.

F. Caselli and W. J. Coleman II. The u.s. structural transformation and regional convergence: A reinterpretation. *Journal of Political Economy*, 109(3):584–616, 2001. doi: 10.1086/321015. URL https://doi.org/10.1086/321015.

Centro Nacional de Memoria Historica. *Justicia y paz. Tierras y territorios en las versiones de los paramilitares*. 2012. ISBN 9789896540821.

Centro Nacional de Memoria Historica. *¡Basta Ya! Colombia: Memorias de Guerra y Dignidad*. 2013. ISBN 9789585760844. URL www.centrodememoriahistorica.gov.co.

J.-P. Chavas. Chapter 5 Structural change in agricultural production: Economics, technology and policy. In *Handbook of Agricultural Economics*, volume 1 of *Agricultural Production*, pages 263–285. Elsevier, 2001.

A. Chayanov. *The Theory of Peasant Economy*. The American Economic Association. Translation series. American Economic Association, 1966.

C. Chen. Untitled land, occupational choice, and agricultural productivity. *American Economic Journal: Macroeconomics*, 9:91–121, 2017. ISSN 19457715. doi: 10.1257/mac.20140171.

P. Collier and S. Dercon. African Agriculture in 50 Years: Smallholders in a Rapidly Changing World? *World Development*, 63:92–101, 2014. doi: 10.1016/j.worlddev.2013.10.001. URL http://dx.doi.org/10.1016/j.worlddev.2013.10.001.

D. A. N. d. E. DANE. Censo nacional agropecuario 2014, avance de resultados. https://www.dane.gov.co/files/CensoAgropecuario/avanceCNA/CNA_agosto_2015_new_present.pdf, 2015. Accessed: 2022–04-29.

D. A. N. d. E. DANE. Metodología General Tercer Censo Nacional Agropecuario, 2016.

D. A. N. d. E. DANE. Sistema de información de precios (sipsa), 2022.

A. De Janvry and E. Sadoulet. Rural Development in Latin America: Relinking Poverty Reduction to Growth. In M. Lipton and J. Van Der Gaag, editors, *Including The Poor*, chapter 11. The World Bank, 1993.

K. Deininger and F. Xia. Quantifying spillover effects from large land-based investment: The case of mozambique. *World Development*, 87(C):227–241, 2016. URL https://EconPapers.repec.org/RePEc:eee:wdevel:v:87:y:2016:i:c:p:227-241.

K. Deininger, D. Byerlee, J. Lindsay, A. Norton, H. Selod, and M. Stickler. Rising Global Interest in Farmland: Can It Yield Sustainable and Equitable Benefits? Technical report, 2011.

K. Deininger, S. Jin, Y. Liu, and S. K. Singh. Can labor market imperfections explain changes in the inverse farm size-productivity relationship? longitudinal evidence from rural india, 2016. URL http://econ.worldbank.org.

Departamendo Nacional de Planeación. Manual Metodológico para La Determinación de La Unidad Agricola Familiar Promedio Municipal. Technical report, 2000. URL https://www.dnp.gov.co/Portals/0/archivos/documentos/DDRS/Publicaciones{_}Estudios/ManualUAF.pdf.

K. Desmet and E. Rossi-Hansberg. Spatial development. *American Economic Review*, 104(4):1211–43, April 2014. doi: 10.1257/aer.104.4.1211. URL https://www.aeaweb.org/articles?id=10.1257/aer.104.4.1211.

R. Dix-Carneiro and B. K. Kovak. Margins of labor market adjustment to trade. *Journal of International Economics*, 117:125–142, 2019. ISSN 0022-1996. doi: https://doi.org/10.1016/j.jinteco.2019.01.005. URL https://www.sciencedirect.com/science/article/pii/S0022199619300078.

F. Eckert and M. Peters. Spatial structural change. Meeting Papers 98, 2018 Society for Economic Dynamics, 2018. URL https://ideas.repec.org/p/red/sed018/98.html.

M. Edel. Determinants of Investments by Colombian Community Action Boards. *The Journal of Developing Areas*, 5(2):207–220, 1971.

M. Eswaran and A. Kotwal. Access to Capital and Agrarian Production Organisation. *The Economic Journal*, 96(382):482–498, 1986.

J.-P. Faguet, F. Sánchez, and M.-J. Villaveces. The perversion of public land distribution by landed elites: Power, inequality and development in colombia. *World Development*, 136:105036, 2020. ISSN 0305-750X. doi: https://doi.org/10.1016/j.worlddev.2020.105036. URL http://www.sciencedirect.com/science/article/pii/S0305750X20301625.

D. Fajardo. *Las Guerras de La Agricultura Colombiana*. 2014.

G. Feder. The relation between farm size and farm productivity: The role of family labor, supervision and credit constraints. *Journal of Development Economics*, 18(2):297–313, 1985. ISSN 0304-3878. doi: https://doi.org/10.1016/0304-3878(85)90059-8. URL https://www.sciencedirect.com/science/article/pii/0304387885900598.

A. Foster and M. Rosenzweig. Economic Development and The Decline of Agricultural Employment. In *Handbook of Development Economics*, volume 4, pages Chapter–47. 2008.

A. D. Foster. Creating Good Employment Opportunities for the Rural Sector. 2011. URL https://www.adb.org/sites/default/files/publication/29119/economics-wp271.pdf.

A. D. Foster and M. R. Rosenzweig. Are there too many farms in the world? labor market transaction costs, machine capacities, and optimal farm size. *Journal of Political Economy*, 130(3):636–680, 2022. doi: 10.1086/717890. URL https://doi.org/10.1086/717890.

M. Gáfaro, A. M. Ibáñez, and D. Zarruk. Equidad y Eficiencia Rural en Colombia: Una Discusión de Políticas Para El Acceso a La Tierra. In A. Montenegro and M. Meléndez, editors, *Equidad y Movilidad Social: Diagnósticos y propuestas para la transformación de la sociedad colombiana*, chapter 11. 2014. URL www.cadena.com.co.

S. Haggblade, P. Hazell, and T. Reardon. *Transforming the Rural Nonfarm Economy: Opportunities and Threats in the Developing World*. International Food Policy Research Institute Series. Johns Hopkins University Press, 2007. ISBN 9780801886645. URL https://books.google.com/books?id=5QNHAwAAQBAJ.

S. Haggblade, P. B. R. Hazell, and T. Reardon. Transforming the rural nonfarm economy: Opportunities and threats in the developing world. Technical report, 2009.

F. Hamann, F. A. Rodríguez, J. A. B. Rojas, M. M. G. González, J. C. M. Vizcaíno, and A. P. P. Olarte. Productividad total de los factores y eficiencia en el uso de los recursos productivos en colombia. *Ensayos Sobre Politica Economica*, 2019:1–54, 2 2019. ISSN 01204483. doi: 10.32468/espe.89.

S. M. Helfand and M. P. Taylor. The inverse relationship between farm size and productivity: Refocusing the debate. *Food Policy*, 99:101977, 2021. ISSN 03069192. doi: 10.1016/j.foodpol.2020.101977. URL https://doi.org/10.1016/j.foodpol.2020.101977.

H. Henderson and A. G. Isaac. Modern value chains and the organization of agrarian production. *American Journal of Agricultural Economics*, 99:379–400, 3 2017. ISSN 14678276. doi: 10.1093/ajae/aaw092.

B. Herrendorf, R. Rogerson, and Ákos Valentinyi. Two perspectives on preferences and structural transformation. *American Economic Review*, 103:2752–2789, 2013. ISSN 00028282. doi: 10.1257/aer.103.7.2752.

B. Herrendorf, R. Rogerson, and Ákos Valentinyi. Chapter 6 - growth and structural transformation. In P. Aghion and S. N. Durlauf, editors, *Handbook of Economic Growth*, volume 2 of *Handbook of Economic Growth*, pages 855–941. Elsevier, 2014. doi: https://doi.org/10.1016/B978-0-444-53540-5.00006-9. URL https://www.sciencedirect.com/science/article/pii/B9780444535405000069.

I. G. A. C. IGAC. *Atlas de la Distribucion de la Propiedad Rural en Colombia*. 2012.

J. Lanjouw and P. Lanjouw. The rural non-farm sector: issues and evidence from developing countries. *Agricultural Economics*, 26(1):1–23, 2001. URL https://EconPapers.repec.org/RePEc:eee:agecon:v:26:y:2001:i:1:p:1-23.

C. Liao, S. Jung, D. G. Brown, and A. Agrawal. Spatial patterns of large-scale land transactions and their potential socio-environmental outcomes in cambodia, ethiopia, liberia, and peru. *Land Degradation & Development*, 31(10):1241–1251, 2020. doi: https://doi.org/10.1002/ldr.3544. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/ldr.3544.

M. Ma, J. Lin, and R. J. Sexton. The transition from small to large farms in developing economies: A welfare analysis. *American Journal of Agricultural Economics*, 2021. ISSN 14678276. doi: 10.1111/ajae.12195.

A. Machado and R. Suarez. *El mercado de tierras en Colombia : una alternativa viable?* Tercer Mundo Editores : CEGA : IICA, Santafé de Bogotá, Colombia, 1999. ISBN 958601858X.

J. W. Mellor. *Agricultural Development and Economic Transformation Promoting Growth with Poverty Reduction*. Palgrave Macmillan, 2017. URL http://www.springer.com/series/14651.

G. Michaels, F. Rauch, and S. J. Redding. Urbanization and Structural Transformation *. *The Quarterly Journal of Economics*, 127(2):535–586, 03 2012. ISSN 0033-5533. doi: 10.1093/qje/qjs003. URL https://doi.org/10.1093/qje/qjs003.

D. Nagy. City location and economic development. Technical report, 2016.

K. Ortiz-Becerra. Land consolidation and structural transformation in rural economies: Evidence from colombia, 2022.

A. Otero-Cortés. El mercado laboral rural en Colombia, 2010-2019. Documentos de trabajo sobre Economía Regional y Urbana 281, Banco de la Republica de Colombia, Nov. 2019. URL https://ideas.repec.org/p/bdr/region/281.html.

J. J. Perfetti, D. Escobar, F. Castro, B. Cuervo, M. Rodríguez, J. I. Vargas, S. M. Bustos, and S. Cortés Acosta. Costos de Producción de Doce Productos Agropecuarios. Technical report, 2012.

M.-C. Piniero. Globalization and Industrialization of Agriculture: Impacts on Rural Chocontá, Colombia. *Luna Azul*, 43:468–498, may 2016. ISSN 1909-2474. doi: 10.17151/luaz.2016.43.20.

L. Putterman. A modified collective agriculture in rural growth-with-equity: Reconsidering the private, unimodal solution. *World Development*, 11(2):77–100, February 1983.

G. Ranis and F. Stewart. Rural nonagricultural activities in development: Theory and application. *Journal of Development Economics*, 40:75–101, 1993. ISSN 0304-3878. doi: 10.1016/0304-3878(93)90105-V. URL http://www.sciencedirect.com/science/article/pii/030438789390105Vhttp://files/526/RanisandStewart-1993-RuralnonagriculturalactivitiesindevelopmentT.pdfhttp://files/527/030438789390105V.html.

O. Robineau, M. Châtelet, C.-T. Soulard, I. Michel-Dounias, and J. Posner. Integrating Farming and Páramo Conservation: A Case Study From Colombia. *Mountain Research and Development*, 30(3):212–221, 2010. ISSN 0276-4741. doi: 10.1659/mrd-journal-d-10-00048.1.

C. C. D. R. R. Santaeulàlia-Llopis. The effects of land markets on resource allocation and agricultural productivity, 2021. URL http://www.nber.org/papers/w24034.

A. Sen. Market failure and control of labour power: towards an explanation of 'structure 9 and change in Indian agriculture. Part 1. *Cambridge Journal of Economics*, 5:201–228, 1981. URL https://academic.oup.com/cje/article-abstract/5/3/201/1686512.

A. K. Sen. Peasants and dualism with or without surplus labor. *Journal of Political Economy*, 74(5):425–450, 1966. ISSN 00223808, 1537534X. URL http://www.jstor.org/stable/1829592.

M. H. Sial and M. R. Carter. Financial market efficiency in an agrarian economy: Microeconometric analysis of the pakistani punjab. *The Journal of Development Studies*, 32(5):771–798, 1996. doi: 10.1080/00220389608422439. URL https://doi.org/10.1080/00220389608422439.

S. Sotelo. Domestic trade frictions and agriculture. *Journal of Political Economy*, 128, 2020.

J. E. Taylor and G. A. Dyer. Migration and the sending economy: A disaggregated rural economy-wide analysis. *The Journal of Development Studies*, 45(6):966–989, 2009. doi: 10.1080/00220380802265553. URL https://doi.org/10.1080/00220380802265553.

J. Tenjo and C. Jaimes. Mercado laboral en el sector rural colombiano. informe para la misión para la transformación del campo., 2015.

2A Data Sources and Variables

National Agricultural Census: this census, conducted in 2013, collects information on agricultural production and practices in the country. This census covers the scattered rural area from all the municipalities and includes all the agricultural production units (UPAs) regardless of titling and tenure regime.7 From the 2.4 million agricultural units that are part of this census, 86% are held by private individuals and companies, and 14% are held by communities under collective rights. For each production unit, this instrument collects data on location, size, land use, number of workers, machinery possession, credit access, and livestock inventory. In addition, it collects information on production, final destination, and agricultural practices for each crop produced. This survey also contains data on the number of households that occupy the UPA, the dwellings' features, and the sociodemographic characteristics of the habitual residents.

I focus on the set of agricultural production units held by private individuals and companies, where the concept of land consolidation is relevant. Thus, I exclude those production units located in indigenous and afro-Colombian communities, which have collective rights to their land. Besides, I exclude the farms that are located in protected areas according to the Agricultural Rural Planning Unit (UPRA), since agriculture is not meant to be the main land vocation in these locations. Finally, following Hamann et al. (2019), I also exclude those production units below 100 square meters and the farms above the average size in three standard deviations or more. My sample accounts for seventy-four percent of the total farmland and has an average area size of 17.2 hectares.

The following are the definitions of the main variables I use in this analysis:

Size of agricultural production unit: area of all the plots of the production unit, in hectares.

Number of agricultural workers: number of permanent workers in the production unit doing agricultural activates over the last thirty days plus (including main producer and family members) plus number of additional "labor days" paid over the same time period. To make these two measures comparable, I follow Hamann et al. (2019) and assume that a permanent worker corresponds to six "labor days" in a week and 24 "labor days" in a month.

Machinery possession: whether there is at least one machinery in the farm for the development of agricultural activities or not.

Main crop: to identify the main crop in the agricultural unit , I focus on the crop that is produced in the largest plot. If the largest plot produces more than one crop, however, I focus on the crop that has had the largest production in the last couple of years.[49]

[49]The census collects information on production for the years 2012 and 2013. Focusing on production over a span of two years allows me to avoid comparability issues due to the seasonality of crops.

National Household Surveys: I use the National Household Survey of 2016 to obtain information on the allocation of employment across economic sectors. This survey collects quarterly data on employment and socio-demographic characteristics for a representative sample of the population. This information is also representative at the level of rural and urban areas.

To classify employment by sector, I use information on the industry in which individuals worked in the last two weeks. I define agricultural employment as those activities that make an intensive use of land for two main purposes: i) the production of food crops and raw materials such as cotton and tobacco, and ii) the raising of livestock. This category excludes those activities related to the processing and elaboration of food and beverages, and the transformation of raw materials since they are not characterized by an intensive used of land. Conversely, I define non-farm employment as all other activities that do not belong to agriculture, which implies that this sector is composed of several industries.

National Households' Budget Survey: I use aggregate data from the National Households' Budget Survey of 2017 to obtain information on the final expenditure share of rural households across sectors. I use this data to calculate the expenditure shares in agricultural goods and nonfarm tradables. On one hand, the share in agricultural goods refers to the expenditures in food and non-alcoholic drinks. On the other, the share in nonfarm tradables refers to expenditures in home furniture, clothing, footwear, alcoholic drinks, and restaurants and hotels.

Study on the costs of agricultural production conducted by Perfetti et al. (2012): This study, conducted in 2012, collects data on the structure of production costs for twelve agricultural products in Colombia (rice, maize, potatoes, flowers, oil palm, cocoa, coffee, bananas, sugar cane, milk, livestock and chicken). This information was collected using close to 1,600 surveys to producers, input sellers, and technical assistants in agricultural production. For most crops, these data are representative of the agricultural setting in regions with the largest participation in production.

From these twelve crops, four have information on the structure of production costs by farm size categories: oil palm, potatoes, maize, and rice. Farms with a size below to one family farm unit are considered small, while farms with one family farm unit or more are considered large.[50] In this analysis, I use data from these four crops to calculate the ratio of farm equipment cost per hectare between small and large farms. These costs are associated with the use of motorized equipment to conduct tasks across all stages of production. According to details from this study, some of the stages with a higher prevalence of mechanization are harvest and land preparation.

[50]The family farm unit is a policy instrument representing the minimum plot size needed to generate an income surplus given the agro-ecological conditions of the plot's location. See Ortiz-Becerra (2022) for more details.

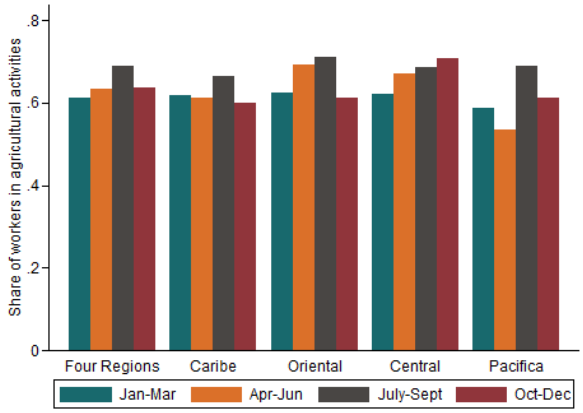
2B Context and Patterns: Additional Figures and Tables

Table 2A: Top Twenty Crops by Sown Area

Crop	Type	Total Area (%)	Total Farms (%)
Coffee	Perennial	10.63	45.07
Plantain	Perennial	9.07	28.58
Rice	Seasonal	7.21	4.15
Yellow corn	Seasonal	6.86	10.95
African Palm	Perennial	6.02	1.93
Yucca	Seasonal	5.35	15.18
White Corn	Seasonal	4.22	7.99
Sugar Cane (panela)	Perennial	3.38	15.72
Potatoes	Seasonal	3.34	3.62
Sugar Cane (refined sugar)	Perennial	3.06	1.28
Cocoa (grain)	Perennial	2.13	8.73
Pineapple	Perennial	1.47	2.46
Banana	Perennial	1.47	4.49
Kidney Bean	Seasonal	1.24	4.88
Avocado	Perennial	1.05	4.10
Arracacha	Seasonal	0.90	2.80
Ñame	Seasonal	0.84	2.21
Orange	Perennial	0.66	2.49
Banana for exports	Perennial	0.64	0.20
Cotton	Seasonal	0.60	0.68

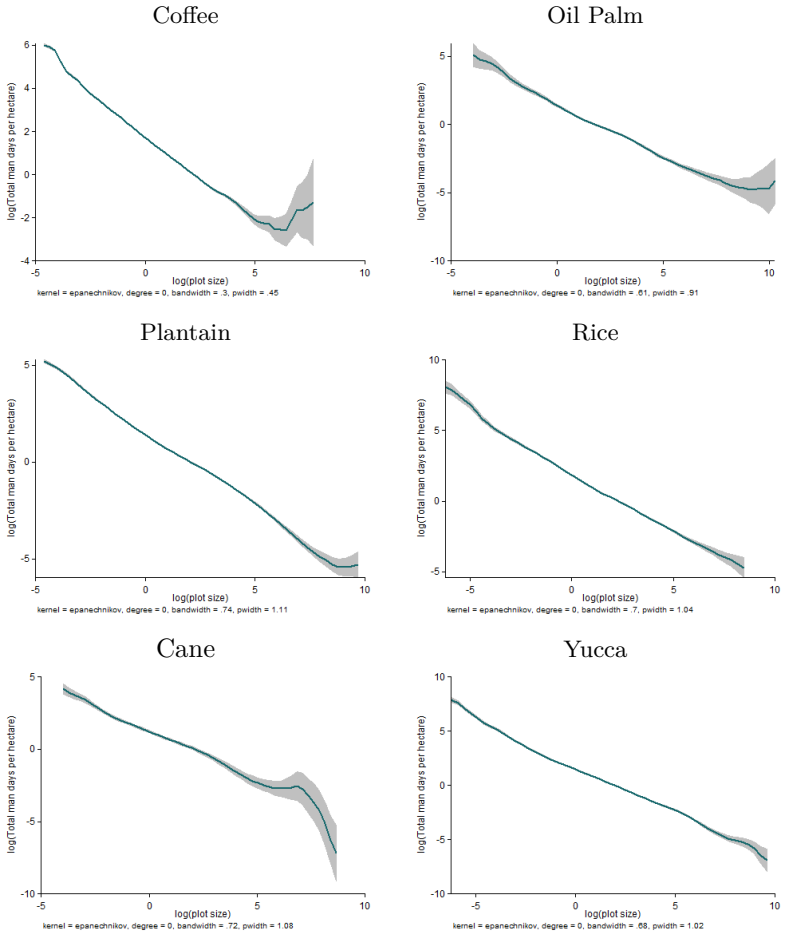
Notes: This table presents the participation of the top twenty main crops in the sown area and the number of farms that produce them. These 20 crops account for 70% of planted area in the country. Since a farmer can plant more than one crop, the shares in the third column do not necessarily add up to one.

Figure 2A: Rural Farm Employment by Quarter and Region



Notes: This figure presents the proportion of agricultural workers in rural areas by quarter and region using data from the national household surveys.

Figure 2B: Labor Intensity and Farm Size By Type of Crop



Notes: Own illustrations using data from the 2014 National Agricultural Census.

2C Model Derivations and Proofs

2C.1 Optimal Input and Consumption Demands

Agricultural production: Since this economy has no active markets for land, only individuals with a positive endowment of hectares produce the agricultural crop. These individuals face a vector of prices $[p_a, p_n, p_c, w, r]$ and choose labor and capital to maximize their profits:

$$\max_{\ell_a, k_a} \quad \pi_a = p_a h_a^{\gamma} [\eta \ell_a^{\rho} + (1-\eta) k_a^{\rho}]^{\frac{(1-\gamma)}{\rho}} - w\ell_a - R(h_a)k_a$$

where $R(h_a) = r(1 + \frac{1}{h_a^{\nu}})$ is the effective price of capital that depends on the benchmark rate r.

The solution to this maximization is characterized by the following first order conditions:

$$(1): \quad \gamma h_a^{(\gamma-1)} [\eta \ell_a^{\rho} + (1-\eta) k_a^{\rho}]^{\frac{(1-\gamma)}{\rho}} = \mu$$

$$(2): \quad (1-\gamma) h_a^{\gamma} [\eta \ell^{\rho} + (1-\eta) k_a^{\rho}]^{\frac{(1-\gamma-\rho)}{\rho}} \eta \ell_a^{\rho-1} = w$$

$$(3): \quad (1-\gamma) h_a^{\gamma} [\eta \ell_a^{\rho} + (1-\eta) k^{\rho}]^{\frac{(1-\gamma-rho)}{\rho}} (1-\eta) k_a^{\rho-1} = r(1 + \tfrac{1}{h_a^{\nu}})$$

$$(4): \quad \mu h = 0$$

where μ is the shadow price of land. Thus, since the value of the marginal productivity of land is positive for any finite farm size, $\mu > 0$, and the optimal input demands are:

$$h^* = h_e$$

$$\ell_a^* = h_e \left[\frac{(1-\gamma)\eta p_a}{w} \right]^{\frac{1}{\gamma}} \left[(1-\eta) \left(\frac{(1-\eta)}{\eta} \frac{w}{R(h_e)} \right)^{\frac{\rho}{(1-\rho)}} + \eta \right]^{\frac{(1-\gamma-\rho)}{\gamma\rho}}$$

$$k_a^* = h_e \left[\frac{(1-\gamma)(1-\eta) p_a}{R(h_e)} \right]^{\frac{1}{\gamma}} \left[\eta \left(\frac{\eta}{(1-\eta)} \frac{R(h_e)}{w} \right)^{\frac{\rho}{(1-\rho)}} + (1-\eta) \right]^{\frac{(1-\gamma-\rho)}{\gamma\rho}}$$

where h_e is the land endowment of individual i.

Local nonfarm production: The nonfarm good is produced by a stand-in firm using a CRS technology with labor and capital

$$\max_{L_n, K_n} \quad \pi_n = p_n L_n^{\alpha} K_n^{1-\alpha} - wL_n - rK_n$$

Thus, profits equal zero and the conditional input demands are given by:

$$L_n^* = \overline{Q}_n \left(\frac{\alpha}{(1-\alpha)} \frac{r}{w} \right)^{(1-\alpha)}$$

$$K_n^* = \overline{Q}_n \left(\frac{(1-\alpha)}{\alpha} \frac{w}{r} \right)^{\alpha}$$

Consumption: All individuals in this economy maximize utility conditional on total net income Y:

$$\begin{aligned} \max_{c_a, c_n, c_c} \quad & u(c_a, c_n, c_c) = \omega_a log(c_a - \bar{c_a}) + \omega_n log(c_n) + \omega_c log(c_c + \bar{c_c}) \\ \text{s.t.} \quad & p_a c_a + p_n c_n + p_c c_c \leq Y \\ & c_a, c_n > 0 \quad ; \quad \bar{c_a}, \bar{c_c} > 0; \quad c_c \geq 0 \end{aligned}$$

The solution to this maximization is characterized by the following first order conditions:

$$(1): \quad \frac{\omega_a}{c_a - \bar{c_a}} - \lambda p_a = 0$$

$$(2): \quad \frac{\omega_n}{c_n} - \lambda p_n = 0$$

$$(3): \quad \frac{\omega_c}{c_c + \bar{c_c}} - \lambda p_c + \mu_c = 0$$

$$(4): \quad \mu_c c_c = 0$$

$$(5): \quad Y - p_a c_a - p_n c_n - p_c c_c = 0$$

where λ is the multiplier of the budget constraint and μ_c is the shadow price of the city good.

These preferences admit two potential consumption regimes. If $\mu_c^* > 0$ (i.e., corner solution), individuals do not consume the city good and the optimal consumption demands are:

$$\begin{aligned} c_a^* &= \frac{Y\omega_a + p_a \bar{c_a} \omega_n}{p_a(\omega_a + \omega_n)} \\ c_n^* &= \frac{\omega_n}{p_n}\left(\frac{Y - p_a \bar{c_a}}{\omega_a + \omega_n}\right) \\ c_c^* &= 0 \end{aligned}$$

If $\mu_c^* = 0$ (i.e., interior solution), on the other hand, individuals consume the city good and the consumption demands are given by:

$$\begin{aligned} c_a^* &= \bar{c_a} + \frac{\omega_a}{p_a}(Y - p_a \bar{c_a} + p_c \bar{c_c}) \\ c_n^* &= \frac{\omega_n}{p_n}(Y - p_a \bar{c_a} + p_c \bar{c_c}) \\ c_c^* &= \frac{\omega_c}{p_c}(Y - p_a \bar{c_a} + p_c \bar{c_c}) - \bar{c_c} \end{aligned}$$

Note that these preferences are determined for all individuals as long as $Y > p_a \bar{c_a}$. Total net income Y is heterogeneous across individuals and depends on production and endowments. For farmers, total income equals the sum of agricultural profits and value of labor endowment ($Y = \pi_a^* + w\ell_e$). For landless individuals, on the other hand, $Y = w\ell_e$.

2C.2 Proofs

Proposition 1

Consider the labor-to-land relationship:

$$\frac{\ell^*}{h^*} = \left[\frac{(1-\gamma)\eta^{\frac{(1-\gamma)}{\rho}}\left[1+\left(\frac{1-\eta}{\eta}\right)^{\frac{1}{(1-\rho)}}\left(\frac{w}{R}\right)^{\frac{\rho}{(1-\rho)}}\right]^{\frac{1-\gamma-\rho}{\rho}}}{w}\right]^{\frac{1}{\gamma}}$$

where $R = r(1+\frac{1}{h^\nu})$. Let $\Omega_1 = \left(\frac{\ell^*}{h^*}\right)^\gamma > 0$ and $\Omega_2 = \left[1+\left(\frac{1-\eta}{\eta}\right)^{\frac{1}{(1-\rho)}}\left(\frac{w}{R}\right)^{\frac{\rho}{(1-\rho)}}\right] > 0$.

The change in labor intensity when the plot size increases in one hectares is given by:

$$\frac{1}{\gamma}\Omega_1^{\frac{1-\gamma}{\gamma}}\Omega_2^{\frac{1-\gamma-2\rho}{\rho}}\frac{\rho}{1-\rho}\Big(\frac{wh^\nu}{r(t^\nu+1)}\Big)^{\frac{(2\rho-1)}{(1-\rho)}}\frac{r\nu wh^{(\nu-1)}}{(rh^\nu+r)^2}$$

Thus, since $\nu > 0$, this derivative is negative if and only if $(1-\gamma) < \rho$.

Proposition 2:

I first show that there exists an income threshold $\tilde{Y}$ at which individuals start purchasing the city good. Consider the optimal consumption demand for city goods in the case of interior solution:

$$c_c^* = \frac{\omega_c(Y + p_c\bar{c_c} - p_a\bar{c_a})}{p_c} - \bar{c_c}$$

Thus, for consumption to be positive, it must be that the income of the individual is at least as large as the following threshold,

$$Y > \underbrace{p_c\bar{c_c}\Big(\frac{1-\omega_c}{\omega_c}\Big) + p_a\bar{c_a}}_{\tilde{Y}}$$

This threshold $\tilde{Y}$ is always positive since $0 < \omega_c < 1$. Individuals with an income equal or below this threshold set $c_c^* = 0$ and have an utility of consumption determined by the non-homothetic parameter $\bar{c_c}$. If total labor endowment is greater than this threshold (i.e., $w\ell_e > \tilde{Y}$), there is only one consumption regime and everyone in the economy consumes the city good.

Now, to illustrate the hump-shaped relationship between expenditure share in local nonfarm consumption and income, let $w\ell_e \leq \tilde{Y}$ such that both consumption regimes exist. The expenditure share when individuals do not consume the city good is $\frac{\omega_n(Y-p_a\bar{c_a})}{Y(\omega_a+\omega_n)}$, and $\frac{\omega_n(Y-p_a\bar{c_a}+p_c\bar{c_c})}{Y}$ when individuals consume the city good.

Taking the derivatives of these shares with respect to income, we obtain $\frac{\omega_n p_a \bar{c_a}}{Y^2(\omega_a+\omega_n)}$ and $\frac{\omega_n}{Y^2}(p_a\bar{c_a} - p_c\bar{c_c})$ in the first and second regime, respectively. Thus, if $p_a\bar{c_a} < p_c\bar{c_c}$, the expenditure share in the second regime declines with income, leading to the hump-shaped relationship between income and the budget share allocated to local nonfarm consumption.

2C.3 Identification of $p_c\bar{c_c}$

The consumption of city goods is determined by two main parameters of interest: p_c and $\bar{c_c}$. However, since I do not observe information on prices in the data, I cannot construct a relevant moment to calibrate p_c. This section presents a useful transformation that allows me to jointly identify $p_c\bar{c_c}$ without affecting the demands for c_a and c_n.

Consider the non-homothetic preferences:

$$u(c) = \omega_a log(c_a - \bar{c_a}) + \omega_n log(c_n) + \omega_f log(c_c + \bar{c_c})$$

Since prices are the same for all agents, expenditure in consumption is proportional to quantity consumed. This property implies that I can rewrite the utility component of the city good (c_c) as a function of expenditure, and thus, multiplying and dividing by p_c inside the logarithmic function I get:

$$u(c) = \omega_a log(c_a - \bar{c_a}) + \omega_n log(c_n) + log\left(\frac{p_f c_f + p_c\bar{c_c}}{p_f}\right)$$

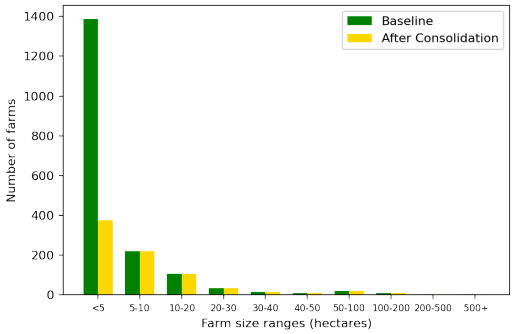
$$u(c) = \omega_a log(c_a - \bar{c_a}) + \omega_n log(c_n) + log(p_c c_c + p_c\bar{c_c}) - log(p_c)$$

Based on these preferences, the solution to the consumer problem yields optimal demands for food and nontradables (c_a^*, c_n^*) and optimal expenditure for the city good $(p_c c_c^*)$. This transformation allows me to jointly identify $p_c\bar{c_c}$ using one moment for both parameters. Since $log(p_c)$ is the same for everyone (i.e., constant), the optimal demands for c_a and c_n do not change.

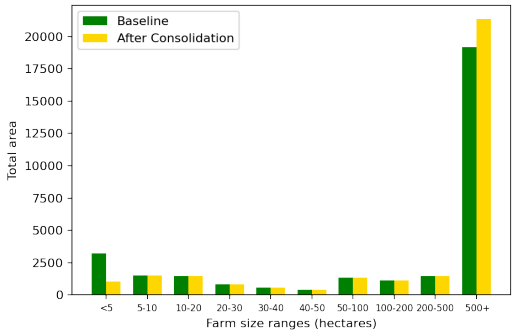
2D Policy Analysis: Additional Figures and Tables

Figure 2C: Farm Size Distribution Before and After Merging Smallest Farms

Panel A. Farm size Distribution



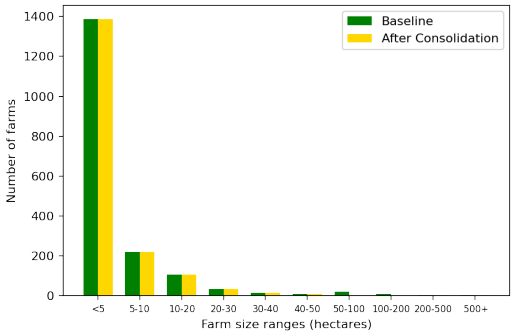
Panel B. Land Distribution



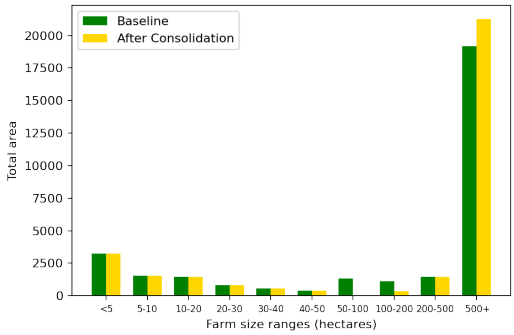
Notes: These figures present the distribution of farms and land before and after the experiment that merges the smallest farms described in Section 2.6.2. Panel A presents the number of farms across size ranges and Panel B presents the total area in each size range.

Figure 2D: Farm Size Distribution Before and After Merging Middle-Sized Farms

Panel A. Farm Size Distribution



Panel B. Land Distribution



Notes: These figures present the distribution of farms and land before and after the experiment that merges middle-sized farms farms described in Section 2.6.2. Panel A presents the number of farms across size ranges and Panel B presents the total area in each size range.

ACKNOWLEDGMENTS

I am deeply grateful to Ashish Shenoy, Michael Carter, Dalia Ghanem, and Steve Boucher for providing guidance and advising on this project and throughout my degree. Many thanks to Douglas Gollin, Diego Restuccia, Jason Shogren, Gaurav Chiplunkar, Esteban Quiñonez, Ameet Morjaria, Debi Mohapatra, John Morehouse, Eleanor Wiseman, Fabio Sanchez, Margarita Gafaro, Jessica Rudder, Aleksander Michuda, and seminar audiences at MWIEDC, UC Davis, the 1st Workshop in Applied Microeconomics, and the LACEA-RIDGE Job Market Showcase for helpful discussions and valuable feedback. I also thank Olam International, its team of in-country operators, and the farmers in rural regions of Antioquia, Huila, Casanare, and Meta for their time during my fieldwork visits throughout the summer and fall of 2019. I acknowledge the support from the UC Davis Provost's Dissertation Fellowship, the Bert and Nell Krantz Fellowship, and the Henry A. Jastro Graduate Research Award.

Chapter 3

Testing Attrition Bias in Field Experiments

Dalia Ghanem, Sarojini Hirshleifer, and Karen Ortiz-Becerra

3.1 Introduction

Randomized control trials (RCTs) are an increasingly important tool of applied economics since, when properly designed and implemented, they can produce internally valid estimates of causal impact.[1] Non-response on outcome measures at endline, however, is an unavoidable threat to the internal validity of many carefully implemented trials. Long-distance migration can make it prohibitively expensive to follow members of an evaluation sample. Conflict, intimidation or natural disasters sometimes make it unsafe to collect complete response data. In high-income countries, survey response rates are often low and may be declining.[2] The recent, increased focus on the long-term impacts of interventions has also made non-response especially relevant. Thus, researchers often face the question: How much of a threat is attrition to the internal validity of a given study?

In this paper, we approach attrition in field experiments with baseline data as an identification problem in a nonseparable panel model. We focus on two identification questions generated by attrition in this setting. First, does the difference in mean outcomes between treatment and control respondents identify the average treatment effect for the respondent subpopulation (ATE-R)? Second, is this estimand equal

[1] Since in the economics literature the term "field experiment" generally refers to a randomized controlled trial, we use the two terms interchangeably in this paper. We do not consider "artefactual" field experiments, also known as "lab experiments in the field," since attrition is often not relevant to such experiments.

[2] See, for example, Meyer et al. (2015) and Barrett et al. (2014).

to the average treatment effect for the study population (ATE)?[3] To answer these questions, we examine the testable implications of the relevant identifying assumptions and propose procedures to test them. Our results provide insights that are relevant to current empirical practice.

We first conduct a systematic review of 96 recent field experiments with baseline outcome data in order to document attrition rates and understand how authors test for attrition bias. Attrition and attrition tests are both common in published field experiments. Although we find wide variation in the choice and implementation of attrition tests in the literature, we are able to identify two main types: (i) a *differential attrition rate test* that determines if attrition rates are different across treatment and control groups, and (ii) a *selective attrition test* that attempts to determine if the mean of baseline observable characteristics differs across the treatment and control groups conditional on response status. While authors report a differential attrition rate test for 79% of field experiments, they report a selective attrition test only 60% of the time. In addition, for a substantial minority of field experiments (36%), authors conduct a *determinants of attrition test* for differences in the distributions of respondents and attritors.

Next, we present a formal treatment of attrition in field experiments with baseline outcome data. Specifically, we establish the identifying assumptions in the presence of attrition for two cases that are likely to be of interest to the researcher. For the first case, in which the researcher's objective is internal validity for the respondent subpopulation (IV-R), the identifying assumption is random assignment conditional on response status (IV-R assumption). This implies that the difference in the mean outcome across the treatment and control respondents identifies the ATE-R, a local average treatment effect for the respondents.[4] In the second case, where internal validity for the study population (IV-P) is of interest, the identifying assumption is that the unobservables that affect response and outcome are independent in addition to the initial random assignment of the treatment (IV-P assumption). If this identifying assumption holds, the ATE for the study population is identified. This second case is especially relevant in settings where the study population is representative of a larger population.

We then derive testable restrictions for each of the above identifying assumptions. If treatment effects for the respondents are the researchers' object of interest, they can implement a test of the IV-R assumption. The null hypothesis of the IV-R test consists of two equality restrictions on the baseline outcome distribution; specifically, for treatment and control respondents as well as treatment and control attritors. Alternatively, if the researchers are interested in treatment effects for the study population, they can test the restriction of the IV-P assumption. The hypothesis of the IV-P test is the equality of the baseline outcome distribution across all four treatment/response subgroups. We show that these testable restrictions are sharp, meaning that they

[3]We refer to the population selected for the evaluation as the study population.

[4]For brevity, we use a "difference in means" to refer to a "difference in population means". To distinguish it from its sample analogue, we refer to the latter as a "difference in sample means".

are the strongest implications that we can test given the available data.[5] We also propose randomization procedures to test the sharp distributional restrictions implied by each identifying assumption as well as regression-based procedures to test their mean counterparts.

In a motivating example, we apply our proposed attrition tests to the randomized evaluation of the *Progresa* program in which the study population is representative of a broader population of interest. We focus on two main outcomes, school enrollment and adult employment. The IV-R test does not reject for either of these outcomes, which is promising for the identification of the treatment effects for the respondent subpopulation. Interestingly, the IV-P test rejects for school enrollment, but it does not reject for adult employment. Thus, for school enrollment only, we reject the internal validity of its treatment effects for the study population. This application illustrates that attrition can have differential implications for the interpretation of treatment effects for different outcomes, even those collected in the same survey. An important takeaway from our analysis is that researchers should consider an outcome-specific approach to testing for attrition bias.

Given their relevance to current empirical practice, we also provide a formal treatment of the differential attrition rate test and the use of covariates. In order to understand the role of differential attrition rates for internal validity, we apply the framework of partial compliance from the local average treatment effect (LATE) literature to potential response.[6] We demonstrate that even though equal attrition rates are sufficient for IV-R under additional assumptions, they are not a necessary condition for internal validity in general. We illustrate using an analytical example and simulations that it is possible to have differences in attrition rates across treatment and control groups while internal validity holds not only for the respondent subpopulation but also the study population. Next, we examine the use of covariates in testing the IV-R or IV-P assumption, which is useful for settings where data on the outcome is not available at baseline. We note two types of covariates that may be included: (i) determinants of the outcome, and (ii) "proxy" variables which are determined by the same variables as the outcome in question. We caution that using covariates that do not fulfill either of these criteria can lead to a false rejection of the IV-R or IV-P assumption.

Finally, we illustrate the empirical relevance of our results by applying our tests to five published field experiments with high attrition rates.[7] A particularly notable result is that, for two-thirds of the outcomes, we neither reject the IV-R nor the IV-P assumption, which ensures the identification of treatment effects for the study population. This is promising for field experiments where the study population is of interest.

[5]Sharp testable restrictions are the restrictions for which there are the smallest possible set of cases such that the testable restriction holds even though the identifying assumption does not. The concept of sharpness of testable restrictions was previously developed and applied in Kitagawa (2015), Hsu et al. (2019), and Mourifié and Wan (2017).

[6]See the foundational work in the LATE literature (Imbens and Angrist, 1994; Angrist et al., 1996).

[7]We choose the five published field experiments from our review that have the highest attrition rates subject to data availability.

For the remaining outcomes, however, our tests reject the IV-P but not the IV-R assumption. In other words, for those outcomes, the researcher would reject the internal validity of the corresponding treatment effect for the study population, but would not reject the assumption that ensures the internal validity of the treatment effect for the respondent subpopulation. When we consider the authors' attrition tests, we find heterogeneity in the choice of tests as well as their implementation, consistent with the findings from our review. Furthermore, our empirical results support the limitations of the differential attrition rate test highlighted by the theoretical analysis. For about one-quarter of the outcomes, our test results are consistent with the conditions under which this test would not control size as a test of internal validity.

This paper has several implications for current empirical practice. First, our theoretical and empirical results imply that the most widely used test in the literature, the differential attrition rate test, may overreject internal validity in practice. The second most widely used test, the selective attrition test, is implemented using a variety of approaches. Most such tests constitute IV-R tests, although those typically use respondents only. Our theoretical results indicate, however, that the implication of the relevant identifying assumption is a joint test that uses all of the available information in the baseline data, and thus includes both respondents and attritors. In addition, while the majority of testing procedures pertain to IV-R and not IV-P, the use of determinants of attrition tests suggests that some researchers may be interested in implications of the estimated treatment effects for the study population. Finally, we note that authors do not typically correct for multiple hypothesis testing in the implementation of selective attrition tests, even when these tests are performed on a non-trivial number of baseline variables. This is another possible source of overrejection of internal validity in the literature. More generally, this paper highlights the importance of understanding the implications of attrition for a broader population when interpreting field experiment results for policy.[8]

This paper contributes to a growing literature that considers methodological questions relevant to field experiments.[9] Given the wide use of attrition tests, we formally examine the testing problem here. Our focus complements a thread in this literature that outlines various approaches to correcting attrition bias in field experiments (Horowitz and Manski, 2000; Lee, 2009; Huber, 2012; Behagel et al., 2015; Millán and Macours, 2019).[10] These corrections build on the vast sample selection literature in econometrics going

[8]External validity can be assessed in a number of ways (see, for example, Andrews and Oster (2019) and Azzam et al. (2018)). In our setting, we note that if IV-R holds but not IV-P, we may be able to draw inference from the local average treatment effect for respondents to a broader population.

[9]Bruhn and McKenzie (2009) compare the performance of different randomization methods; McKenzie (2012) discusses the power trade-offs of the number of follow-up samples in the experimental design; Baird et al. (2018) propose an optimal method to design field experiments in the presence of interference; de Chaisemartin and Behaghel (2018) present how to estimate treatment effects in the context of randomized wait lists; Abadie et al. (2018) propose alternative estimators that reduce the bias resulting from endogenous stratification in field experiments; Muralidharan et al. (2019) examine empirical practice in analyzing experiment with factorial design and analyze the trade-off between power and correct inference in this setting; Kasy and Sautmann (2020) propose a treatment assignment algorithm to choose the best among a set of policies at the end of an experiment; Vazquez-Bare (2020) examines the identification and estimation of spillover effects in randomized experiments.

[10]Other work considers corrections for settings with sample selection and noncompliance. Chen and Flores (2015) rely on monotonicity restrictions to construct bounds for average treatment effects in the presence of partial compliance and sample

back to Heckman (1976, 1979).[11] While the latter literature is broadly concerned with population objects, work that is relevant to program evaluation proposes corrections for objects pertaining to subpopulations (e.g. Lee, 2009; Huber, 2012; Chen and Flores, 2015). Our paper provides tests of identifying assumptions emphasizing the distinction between the (study) population and the respondent subpopulation. Finally, the randomization tests we propose contribute to recent work that examines the potential use of randomization tests in analyzing field experiment data (Young, 2018; Athey and Imbens, 2017; Athey et al., 2018; Bugni et al., 2018).

We also build on other strands of the econometrics literature. Recent work on nonparametric identification in nonseparable panel data models informs our approach (Altonji and Matzkin, 2005; Bester and Hansen, 2009; Chernozhukov et al., 2013; Hoderlein and White, 2012; Ghanem, 2017). Specifically, the identifying assumptions in this paper fall under the nonparametric correlated random effects category (Altonji and Matzkin, 2005). Furthermore, we build on the literature on randomization tests for distributional statistics (Dufour, 2006; Dufour et al., 1998).

The paper proceeds as follows. Section 3.2 presents the review of the field experiment literature. Section 3.3 formally presents the identifying assumptions and their sharp testable restrictions. It also includes a formal treatment of differential attrition rates and of the role of covariates in testing internal validity. Section 3.4 presents simulation experiments to illustrate the theoretical results. Section 3.5 presents the results of the empirical application exercise. Section 3.6 concludes. Sections 3A and 3B present the randomization and regression-based procedures, respectively, to test the IV-R and IV-P assumptions for completely, stratified and cluster randomized experiments.

3.2 Attrition in the Field Experiment Literature

We systematically reviewed 93 recent articles published in economics journals that report the results of 96 field experiments. The objective of this review is to understand both the extent to which attrition is observed and the implementation of tests for attrition bias in the literature.[12] Our categorization imposes some structure

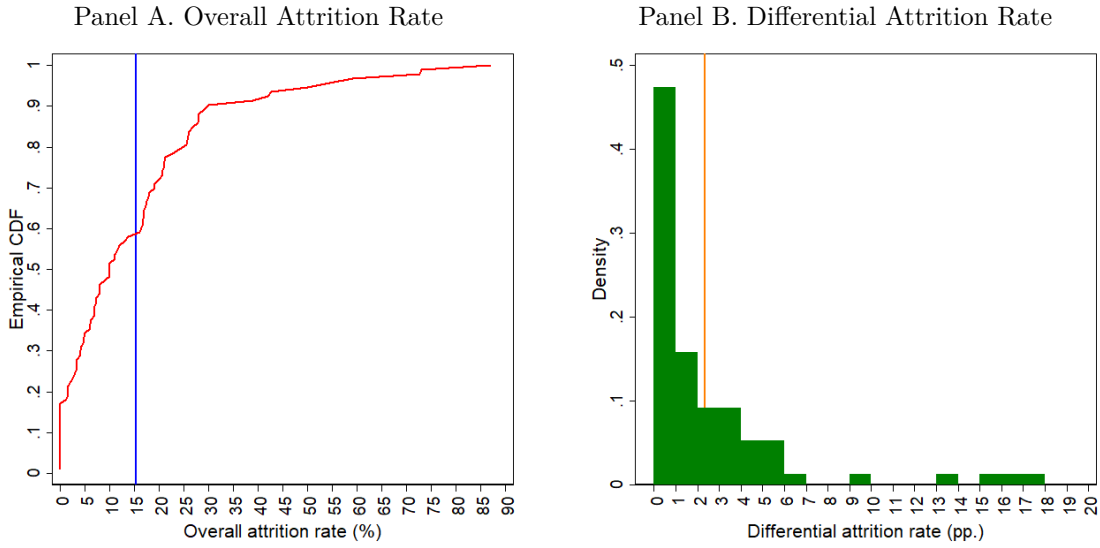
selection. Fricke et al. (2015) consider instrumental variables approaches to address these two identification problems.

[11] Nonparametric Heckman-style corrections have been proposed for linear and nonparametric outcome models (e.g. Ahn and Powell, 1993; Das et al., 2003). Inverse probability weighting (Horvitz and Thompson, 1952; Hirano et al., 2003; Robins et al., 1994) is another important category of corrections for sample selection bias, frequently used in the field experiment literature. Attrition corrections for panel data have also been proposed (e.g. Hausman and Wise, 1979; Wooldridge, 1995; Hirano et al., 2001). Finally, nonparametric bounds is an alternative approach relying on weaker conditions (Horowitz and Manski, 2000; Manski, 2005; Lee, 2009; Kline and Santos, 2013).

[12] We included articles from 2009 to 2015 that were published in the top five journals in economics as well as five highly regarded applied economics journals that commonly publish field experiments: *American Economic Review*, *American Economic Journal: Applied Economics*, *Econometrica*, *Economic Journal*, *Journal of Development Economics*, *Journal of Human Resources*, *Journal of Political Economy*, *Review of Economics and Statistics*, *Review of Economic Studies*, and *Quarterly Journal of Economics*. Section 3D.1 in the online appendix includes additional details on the selection of papers and relevant attrition rates. Section 3I in the online appendix contains a list of all the papers included in the review.

on the variety of different estimation strategies used to test for attrition bias in the literature.[13] In keeping with our panel approach, we focus on field experiments in which the authors had baseline data on at least one main outcome variable.[14]

Figure 3.1: Attrition Rates Relevant to Main Outcomes in Field Experiments



Notes: We report one observation per field experiment. Specifically, the highest attrition rate relevant to a result reported in the abstract of the article. The *Overall* rate is the attrition rate for the full sample, which is composed of the treatment and control groups. The *Differential* rate is the absolute value of the difference in attrition rates across treatment and control groups. The blue (orange) line depicts the average overall (differential) attrition rate in our sample of field experiments. Panel A includes 93 field experiments and Panel B includes 76 experiments since the relevant attrition rates are not reported in some articles.

We review reported overall and differential attrition rates in field experiment papers and find that attrition is common. As depicted in Panel A in Figure 3.1, even though 22% of field experiments have less than 2% attrition overall, the distribution of attrition rates has a long right tail. Specifically, 45% of reviewed field experiments have an attrition rate higher than the average of 15%.[15] Of the experiments that report a differential attrition rate, Panel B in Figure 3.1 illustrates that a majority have little differential attrition for the abstract results: 63% have a differential rate that is less than 2 percentage points, and only 11% have a differential attrition rate that is greater than 5 percentage points.[16]

[13]We identify fifteen estimation strategies used to conduct attrition tests (see Section 3E in the online appendix).

[14]We exclude 64 field experiments that were published during that time period, since they lack baseline data for any outcome mentioned in the abstract. Of those, slightly less than half (44%) are experiments for which the baseline outcome is the same for everyone by design and hence is not informative (see Section 3D.1 in the online appendix).

[15]To understand the extent of attrition that is relevant to the main outcomes in the paper, we focus on attrition rates that are relevant to outcomes reported in the abstract (i.e. "abstract results"). Most papers report attrition rates at the level of the data source or subsample, rather than at the level of the outcome. Since the number of data sources and/or subsamples that are relevant to the abstract results vary by experiment, we include one attrition rate per field experiment for consistency. Specifically, we report the highest attrition rate relevant to an abstract result. Authors do not in general report attrition rates conditional on baseline response. A noteworthy finding from Table 3B in the online appendix is that attrition rates are higher on average for experiments in high-income countries.

[16]It is possible, however, that these numbers reflect authors' exclusion of results with higher differential attrition rates than those that were reported or published.

We then study how authors test for attrition bias. Notably, attrition tests are widely used in the literature: 92% of field experiments with an attrition rate of at least 1% for an outcome with baseline data conduct at least one attrition test. We first identify two main types of tests that aim to determine the impact of attrition on internal validity: (i) a *differential attrition rate test*, and (ii) a *selective attrition test*. A *differential attrition rate test* determines whether the rates of attrition are statistically significantly different across treatment and control groups. In contrast, a *selective attrition test* aims to determine whether, conditional on being a respondent and/or attritor, the mean of observable characteristics is the same across treatment and control groups. We find that there is no consensus on whether to conduct a differential attrition rate test or a selective attrition test, however (Panel A in Table 3.1). In the field experiments that we reviewed, the differential attrition rate test is substantially more common (79%) than the selective attrition test (60%). In fact, 30% of the articles that conducted a differential attrition rate do not conduct a selective attrition test.[17]

Table 3.1: Distribution of Field Experiments by Attrition Test

Panel A: Differential and Selective Attrition Tests

Proportion of field experiments that conduct:		Selective attrition test		
		No	*Yes*	*Total*
Differential attrition rate test	*No*	10%	10%	21%
	Yes	30%	49%	79%
	Total	40%	60%	100%

Panel B: Types of Selective Attrition Test

Conditional on conducting a selective attrition test:	
Test using respondents and attritors	29%
Test using respondents only	67%
Test using attritors only	4%
Total†	100%

Panel C: Determinants of Attrition Tests

Proportion of field experiments that conduct:	Determinants of attrition test		
	Yes	*No*	*Total*
Differential attrition rate test only	12%	18%	30%
Selective attrition test only	1%	9%	10%
Differential & selective attrition tests	21%	28%	49%
No differential & no selective attrition test	1%	9%	10%
Total	36%	64%	100%

Notes: Panel A and C include 77 field experiments that have an attrition rate of at least 1% for an outcome with baseline data. Panel B includes 46 of those experiments that conducted a selective attrition test (†). For details on the classification of the empirical strategies, see Section 3E in the online appendix.

We further consider if selective attrition tests include both respondents and attritors or if they include

[17] We also consider some potential determinants of the use of selective attrition tests: overall attrition rates, differential rates, year of publication, journal of publication. We do not find any strong correlations given the available data.

either only respondents or only attritors (Panel B in Table 3.1). Conditional on having conducted any type of selective attrition test, authors include both respondents and attritors in only 29% of those field experiments. Instead, authors conduct a selective attrition test on the sample of respondents in most cases (67%). Although our review is limited to experiments in which baseline outcome data is available, covariates are typically included in attrition tests along with the baseline outcome. In particular, 98% of field experiments that report a selective attrition test include more than one baseline variable in that test.[18] A key issue that arises with the inclusion of covariates is how to approach the issue of multiple testing. We find that 75% of the experiments that implement a selective attrition test conduct it on an average of 16 variables, and none of those implement a multiple testing correction (Table 3C in online appendix). Only a minority of authors conduct a joint test across all of the baseline variables included in the test (25%).

Another important aspect of testing for attrition bias is testing for differences in the distributions of respondents and attritors. Such tests can illustrate the implications of the main results of the experiment for the study population. We define a *determinants of attrition test* as a test of whether baseline outcomes and covariates correlate with response status and find that authors conduct such a test in approximately one-third of field experiments (Panel C of Table 3.1). Table 3.1 illustrates that conducting the determinants of attrition test does not have a one-to-one relationship with either conducting a differential attrition rate test or conducting a selective attrition test.[19]

3.3 Testing Attrition Bias Using Baseline Data

This section presents a formal treatment of attrition in field experiments with baseline outcome data. First, we motivate the problem with an example from the *Progresa* evaluation. Then, we present the identifying assumptions in the presence of non-response and show their sharp testable implications when baseline outcome data is available for both completely and stratified randomized experiments. We further examine the role of the widely-used differential attrition rate test and discuss the implications of our theoretical analysis for empirical practice.

[18] Although identifying which variables are outcomes or covariates is beyond the scope of this paper, we note that in 91% of the experiments the selective attrition test includes at least one variable that we can easily identify as a covariate (such as age or gender).

[19] Approximately half of the determinants of attrition tests are conducted using the same regression used to test for differential attrition rates. We categorize this strategy as both types of tests since authors typically interpret both the coefficients on treatment and the baseline covariates.

3.3.1 Motivating Example

To illustrate the problem of attrition in field experiments, we use data collected for the randomized evaluation of *Progresa*, a social program in Mexico that provides cash to eligible poor households on the condition that children attend school and family members visit health centers regularly (Skoufias, 2005). The evaluation of *Progresa* relied on the cluster-level random assignment of 320 localities into the treatment group and 186 localities into the control group. These localities, which constitute the study population, were selected to be representative of a larger population of 6396 eligible localities across seven states in Mexico.[20] The surveys conducted for the experiment include a baseline and three follow-up rounds collected 5, 13, and 18 months after the program began.[21] We examine two outcomes of the evaluation that have been previously studied: (i) current *school enrollment* for children 6 to 16 years old, and (ii) paid *employment* for adults in the last week.

Table 3.2: Summary Statistics for the Outcomes of Interest for *Progresa*

	Full Sample				Respondent Subsample at Follow-up			
Round	N	Control Mean	$T-C$	p-value	Attrition Rate	Control Mean	$T-C$	p-value
Panel A. School Enrollment (6-16 years old)								
Baseline	24353	0.824	0.007	0.455				
Pooled					0.183	0.793	0.046	0.000
1st					0.142	0.814	0.043	0.000
2nd					0.234	0.829	0.046	0.000
3rd					0.174	0.740	0.047	0.000
Panel B. Employment Last Week (18+ years old)								
Baseline	31237	0.471	-0.006	0.546				
Pooled					0.161	0.464	0.014	0.002
1st					0.096	0.460	0.016	0.016
2nd					0.196	0.459	0.009	0.138
3rd					0.192	0.472	0.018	0.001

Notes: T and C refer to treatment and control group, respectively. $T-C$ is the difference in sample means between the treatment and control groups and the p-value is estimated with a regression of outcome on treatment that clusters standard errors at the locality level. The attrition rates reported are conditional on responding to the baseline survey. *Pooled* refers to data from all three follow-ups combined.

In Table 3.2, we report the initial sample size and summary statistics for each outcome by treatment group at baseline and follow-up. The failure to reject the null hypothesis of the equality of means across the treatment and control groups at baseline is suggestive evidence that the randomization of localities into treatment and control was implemented correctly. In the context of treatment randomization and absence of attrition, the difference in a mean outcome across treatment and control groups at follow-up would identify the average treatment effect for the study population.[22] Pooling data from the three follow-up rounds, we

[20]Localities were eligible if they ranked high on an index of deprivation, had access to schools and a clinic, and had a population of 50 to 2500 people. See INSP (2005) for details about the experiment. For this analysis, we use the evaluation panel dataset, which can be found on the official website of the evaluation at https://evaluacion.prospera.gob.mx/es/eval_cuant/p_bases_cuanti.php.

[21]The baseline was collected in October 1997 and the three follow-ups were collected in October 1998, June 1999, and November 1999.

[22]Here we follow our convention of referring to a "difference in population means" as a "difference in means."

would conclude that the impact of *Progesa* on school enrollment (adult employment) is an increase of 4.6 (1.4) percentage points. The attrition rate, however, varies from 10% to 24% depending on the outcome and the follow-up round. These attrition rates raise the question of whether these treatment effect estimates are unbiased for at least one of two objects of interest: (i) the average treatment effect for the respondent subpopulation (ATE-R) or (ii) the average treatment effect for the entire study population (ATE).

In order to understand whether attrition affects the internal validity of this experiment, we inspect the mean baseline outcomes across the four treatment-response subgroups. For the outcome of school enrollment, there are two distinct patterns. First, baseline school enrollment is similar across treatment and control respondents as well as treatment and control attritors. Second, we find meaningful differences when we compare respondents and attritors: baseline school enrollment is around 87% for the respondents and 61% for the attritors in the pooled follow-up sample. Taken together, these two patterns suggest that while the unobservables that affect the outcome are correlated with response, they are still independent of the treatment *within* respondents and *within* attritors. As we formalize in the next section, independence between treatment status and the unobservables that affect the outcome conditional on response status constitutes the identifying assumption of internal validity for the respondents (IV-R assumption). We show that the IV-R assumption implies the identification of treatment effects for the respondent subpopulation and that its testable implication is that the distribution of a baseline outcome is identical across treatment and control respondents as well as treatment and control attritors. Applying this test to school enrollment in Column 7 of Table 3.3, we do not reject the IV-R assumption.[23] If the IV-R assumption does hold for this outcome, then the difference in means across treatment and control respondents at follow-up identifies an average treatment effect for the respondents (ATE-R).

Next, we examine the second outcome, adult employment, as observed at baseline. In contrast to school enrollment, adult employment is similar across all four treatment-response subgroups. This pattern indicates that the unobservables that determine the outcome are independent of treatment and response status. This is consistent with the identifying assumption for internal validity for the study population (the IV-P assumption), which we formally define in the next section. We then show that under random assignment the IV-P assumption implies the identification of treatment effects for the study population and its testable implication is indeed that the distribution of baseline outcome is identical across all four treatment-response subgroups. When we formally test the implication of the IV-P assumption for adult employment, we do not reject it (Column 8 of Table 3.3). Thus, we do not reject the assumption that ensures that the difference

[23]Note that the two outcomes we examine here are binary, so the equality of means is equivalent to a distributional equality. It is worth noting that a multiple testing correction would not change the decisions of any of the tests in our example. For instance, applying the Bonferroni correction for each outcome would yield a significance level for each hypothesis of 0.63% to control a family-wise error rate of 5% across the eight tests we conduct.

Table 3.3: Internal Validity in the Presence of Attrition for *Progresa*

Follow-up Sample	Attrition Rate		Mean Baseline Outcome by Group				Test of IV-R	Test of IV-P
	C (1)	Differential (2)	TR (3)	CR (4)	TA (5)	CA (6)	*p*-value (7)	*p*-value (8)
Panel A. School Enrollment (6-16 years old)								
Pooled	0.187	-0.007	0.878	0.874	0.615	0.605	0.836	0.000
1st	0.150	-0.013	0.875	0.871	0.550	0.554	0.810	0.000
2nd	0.244	-0.017	0.901	0.897	0.590	0.595	0.824	0.000
3rd	0.168	0.009	0.859	0.856	0.697	0.663	0.217	0.000
Panel B. Employment Last Week (18+ years old)								
Pooled	0.157	0.007	0.463	0.468	0.472	0.486	0.698	0.132
1st	0.100	-0.007	0.464	0.471	0.472	0.473	0.825	0.860
2nd	0.195	0.001	0.463	0.465	0.474	0.496	0.566	0.058
3rd	0.175	0.027	0.463	0.469	0.471	0.481	0.769	0.503

Notes: The mean baseline outcomes correspond to the groups of treatment respondents (TR), control respondents (CR), treatment attritors (TA), and control attritors (CA). *Pooled* refers to all the three follow-ups. The tests of internal validity were conducted using the regression tests proposed in Section 3B. All regression tests use clustered standard errors at the locality level.

in mean employment rates between treatment and control respondents at follow-up identifies not only the ATE-R but also the average treatment effect (ATE). For the outcome of school enrollment, however, we do reject the IV-P assumption (Column 8 of Table 3.3), and thus the estimated treatment effect cannot be interpreted as internally valid for the study population. This is consistent with our previous observation that the children that are observed in the follow-up data are substantially different at baseline from those that are not.

Understanding treatment effects for the study population is especially relevant to understanding the impact of large-scale programs such as *Progresa*, where the study population is representative of a larger population. In this type of study, if we do reject the IV-P assumption but not the IV-R assumption for an outcome such as school enrollment, we can still draw inferences about an average treatment effect on a larger population. That average treatment effect, however, is a local average treatment effect for the type of participants for which there would be follow-up data available for a given outcome.

3.3.2 Internal Validity in the Presence of Attrition

In this section, we derive the testable implications of our distributional and mean identifying assumptions. We also present the extension of the results to stratified randomization and heterogeneous treatment effects, formally defined as conditional average treatment effects.

Internal Validity and its Testable Restrictions

In a field experiment with baseline outcome data, we observe individuals $i = 1, \ldots, n$ over two time periods, $t = 0, 1$. We will refer to $t = 0$ as the baseline period, and $t = 1$ as the follow-up period. Individuals are randomly assigned in the baseline period to the treatment and control groups. We use D_{it} to denote

treatment status for individual i in period t, where $D_{it} \in \{0,1\}$.[24] Hence, the treatment and control groups can be characterized by $D_i \equiv (D_{i0}, D_{i1}) = (0,1)$ and $D_i = (0,0)$, respectively. For notational brevity, we let an indicator variable T_i denote the group membership. Specifically, $T_i = 1$ if individual i belongs to the treatment group and $T_i = 0$ if individual i belongs to the control group.

For each period $t = 0, 1$, we observe an outcome Y_{it}, which is determined by the treatment status and a $d_U \times 1$ vector of time-invariant and time-varying variables, $U_{it} \equiv (\alpha_i', \eta_{it}')'$,

$$Y_{it} = \mu_t(D_{it}, U_{it}). \tag{3.3.1}$$

Given this structural function, we can define the potential outcomes $Y_{it}(d) = \mu_t(d, U_{it})$ for $d = 0, 1$. We use structural notation here since it is more common in the panel literature. This notation also allows us to refer to the unobservables that affect the outcome, which play an important role in understanding internal validity questions in our problem. To simplify illustration, we postpone the discussion of covariates to Section 3.3.4.

Consider a properly designed and implemented RCT such that by random assignment the treatment and control groups have the same distribution of unobservables. That is, $(U_{i0}, U_{i1}) \perp T_i$, which can be expressed as $(Y_{i0}(0), Y_{i0}(1), Y_{i1}(0), Y_{i1}(1)) \perp T_i$ using the potential outcomes notation. This implies that the control group provides a valid counterfactual outcome distribution for the treatment group, i.e. $Y_{i1}(0)|T_i = 1 \overset{d}{=} Y_{i1}|T_i = 0$, where $\overset{d}{=}$ denotes the equality in distribution. In this case, any difference in the outcome distribution between treatment and control groups in the follow-up period can be attributed to the treatment. The ATE can be identified as the difference in mean outcomes between the treatment and control group,

$$\underbrace{E[Y_{i1}(1) - Y_{i1}(0)]}_{ATE} = E[Y_{i1}|T_i = 1] - E[Y_{i1}|T_i = 0]. \tag{3.3.2}$$

We now introduce the possibility of attrition in our setting. We assume that all individuals respond in the baseline period ($t = 0$), but there is possibility of non-response in the follow-up period ($t = 1$). Response status in the follow-up period is determined by the following equation,[25]

$$R_i = \xi(T_i, V_i), \tag{3.3.3}$$

where V_i denotes a vector of unobservables that determine response status and potential response can be defined as $R_i(\tau) = R_i(\tau, V_i)$ for $\tau = 0, 1$. If individual i responds, then $R_i = 1$, otherwise it is zero. As a

[24] The extension to the multiple treatment case is in Section 3G of the online appendix.

[25] Since non-response is only allowed in the follow-up period, we omit time subscripts from the response equation for notational convenience.

result, instead of observing the outcome for all individuals in the treatment and control groups at follow-up, we can only observe the outcome for respondents in both groups. Random assignment in the presence of attrition, $(U_{i0}, U_{i1}, V_i) \perp T_i$, does not ensure that comparisons between treatment and control respondents are solely attributable to the treatment, since these comparisons are conditional on being able to observe individuals at follow-up ($R_i = 1$).[26]

Two questions arise in this setting. First, do the control respondents provide an appropriate counterfactual for the treatment respondents, $Y_{i1}|T_i = 0, R_i = 1 \stackrel{d}{=} Y_{i1}(0)|T_i = 1, R_i = 1$? This would imply that we can obtain internally valid estimands for the respondent subpopulation, such as the ATE-R, $E[Y_{i1}(1) - Y_{i1}(0)|R_i = 1]$. Second, do the outcome distributions of treatment and control respondents in the follow-up period identify the potential outcome distribution of the study population with and without the treatment, $Y_{i1}|T_i = \tau, R_i = 1 \stackrel{d}{=} Y_{i1}(\tau)$ for $\tau = 0, 1$? This would imply that we can obtain internally valid estimands for the study population, such as the ATE.

The next proposition provides sufficient conditions to obtain each of the aforementioned equalities as well as their respective sharp testable restrictions. Restrictions are sharp when they are the strongest implications that can be tested given the available data (see Figure 3.4). Part *a* (*b*) of the following proposition refers to the case where we can obtain valid estimands for the respondent subpopulation (study population). The proof of the proposition is given in Section 3C.

Proposition 5. *Assume* $(U_{i0}, U_{i1}, V_i) \perp T_i$.[27]

(a) If $(U_{i0}, U_{i1}) \perp T_i | R_i$ *holds, then*

(i) (Identification) $Y_{i1}|T_i = 0, R_i = 1 \stackrel{d}{=} Y_{i1}(0)|T_i = 1, R_i = 1$

(ii) (Sharp Testable Restriction) $Y_{i0}|T_i = 0, R_i = r \stackrel{d}{=} Y_{i0}|T_i = 1, R_i = r$ *for* $r = 0, 1$.

(b) If $(U_{i0}, U_{i1}) \perp R_i | T_i$ *holds, then*

(i) (Identification) $Y_{i1}|T_i = \tau, R_i = 1 \stackrel{d}{=} Y_{i1}(\tau)$ *for* $\tau = 0, 1$.

(ii) (Sharp Testable Restriction) $Y_{i0}|T_i = \tau, R_i = r \stackrel{d}{=} Y_{i0}$ *for* $\tau = 0, 1$, $r = 0, 1$.

Proposition 1(a) relies on the assumption of random assignment conditional on response status (IV-R assumption). This assumption implies that the outcome distributions of treatment and control *respondents* at endline would have been the same if the treatment status had never been assigned. We refer to this equality (a.i) as *internal validity for the respondent subpopulation* (IV-R). When IV-R holds, the difference

[26] We use a random assignment condition similar to Lee (2009). Using potential outcome and response notation, we can express the random assignment condition as $(Y_{i0}(0), Y_{i0}(1), Y_{i1}(0), Y_{i1}(1), R_i(0), R_i(1)) \perp T_i$ which is similar to Lee (2009).

[27] The random assignment condition can be expressed as $(Y_{i0}(0), Y_{i0}(1), Y_{i1}(0), Y_{i1}(1), R_i(0), R_i(1)) \perp T_i$ in potential outcome and response notation.

in means between treatment and control respondents identifies the ATE-R. IV-R cannot be tested directly, however, since treatment was in fact assigned. Thus, we derive a sharp testable restriction (a.ii) of the IV-R assumption, which exploits the information in the baseline data.[28] This restriction implies that the appropriate attrition test (when the object of interest is the treatment effect on the respondent subpopulation) is a *joint test* of the equality of the baseline outcome distribution between treatment and control respondents as well as treatment and control attritors.[29]

The assumption in Proposition 1(b), under random assignment, implies that treatment and response status are jointly independent of the unobservables in the outcome equation.[30] As a result, in the absence of treatment, all four treatment-response sub-groups would have the same outcome distribution. We refer to this case as *internal validity for the study population* (IV-P) and the assumption in (b) as the IV-P assumption. When IV-P holds, the ATE is identified, and so are quantile and other distributional treatment effects for the study population. The sharp testable restriction of the IV-P assumption under random assignment is given in (b.ii).

Mean Tests of Internal Validity

The vast majority of selective attrition tests implemented in the literature are based on restrictions on the mean of the baseline variables in question. The IV-R and IV-P assumptions we present above ensure the identification of distributional treatment effects in addition to average treatment effects. In some experiments, however, researchers may be solely interested in average treatment effects. Here, we discuss the weaker conditions required to identify these objects and their sharp testable implications. Section 3B presents regression-based tests for these restrictions.

If the researcher is interested in mean impacts for the respondent subpopulation, then the IV-R assumption in Proposition 5(a), while sufficient, is stronger than required. A weaker condition that ensures that the average potential outcome without the treatment is identical for treatment and control respondents as

[28] While it is *theoretically* possible for identification to hold while the testable restriction is violated, it is not an interesting case empirically. If a field experimentalist finds violations of the testable implication of the IV-R (or IV-P) assumption at baseline, it is highly unlikely that they will discount this evidence and argue that identification of the ATE-R (or ATE) remains possible from a simple difference of means between treatment and control respondents.

[29] If IV-R is of interest, a natural question is whether one should simply test the implication of $(U_{i0}, U_{i1}) \perp T_i | R_i = 1$ in lieu of the IV-R assumption ($(U_{i0}, U_{i1}) \perp T_i | R_i$). This would be empirically relevant if it is plausible that $(U_{i0}, U_{i1}) \perp T_i | R_i = 1$ holds while $(U_{i0}, U_{i1}) \perp T_i | R_i = 0$ is violated. Using the subgroups defined by potential response status, we note that a primitive condition for this to hold is $(U_{i0}, U_{i1}) | (R_i(0), R_i(1)) \stackrel{d}{=} (U_{i0}, U_{i1}) | max\{R_i(0), R_i(1)\}$. This condition is not empirically plausible since it implies that the unobservable distribution is the same for always-responders, treatment-only and control-only responders, but different for the never-responders.

[30] This implies *missing-at-random* as defined in Manski (2005). In the cross-sectional setup, the missing-at-random assumption is given by $Y_i | T_i, R_i \stackrel{d}{=} Y_i | T_i$. Manski (2005) establishes that this assumption is not testable in that context. We obtain the testable implications by exploiting the panel structure. It is important to emphasize that this definition of missing-at-random is different from the assumption in Hirano et al. (2001) building on Rubin (1976), which would translate to $Y_{i1} \perp R_i | Y_{i0}, T_i$ in our notation. Finally, while we do not distinguish between observables and unobservables here, it is worth noting that Assumption 3 in Huber (2012) provides a set of conditions that imply the assumption in Proposition 5(b).

well as treatment and control attritors, specifically

$$E[Y_{it}(0)|T_i, R_i] = E[Y_{it}(0)|R_i],\ t = 0, 1, \quad \text{(Mean IV-R Assumption)} \tag{3.3.4}$$

implies the identification of the ATE-R. Its sharp testable implication is the mean version of the testable restriction in Proposition 5(a.ii),

$$E[Y_{i0}|T_i, R_i] = E[Y_{i0}|R_i], \tag{3.3.5}$$

so it also includes testable restrictions on attritors and respondents. We will refer to a test of the mean equality restrictions in (3.3.5) as a mean IV-R test.

Similarly, if the object of interest is the ATE for the study population, then the relevant identifying assumption is

$$E[Y_{it}(d)|T_i, R_i] = E[Y_{it}(d)],\ d = 0, 1,\ t = 0, 1, \qquad \text{(Mean IV-P Assumption)} \tag{3.3.6}$$

which ensures that the average potential outcomes are identical across the four treatment-response subgroups. The sharp testable restriction of this assumption,

$$E[Y_{i0}|T_i, R_i] = E[Y_{i0}], \tag{3.3.7}$$

involves all treatment-response subgroups as its distributional version in Proposition 5(b.ii). We will refer to a test based on (3.3.7) as a mean IV-P test.

Heterogeneous Treatment Effects and Stratified Randomization

In this section, we extend our analysis to discuss heterogeneous treatment effects and stratified randomization. Heterogeneous treatment effects, more formally referred to as conditional average treatment effects (CATE), are of interest in many experiments. Stratified randomization is also common in empirical practice. Sometimes it is a necessity of the design, such as when the study is randomized within roll-out waves or locations. At other times, it is included in the experimental design with the aim of increasing precision and reducing bias of both average and heterogeneous treatment effects. The results in this section are relevant both for stratified randomized experiments and for completely randomized experiments that estimate heterogeneous treatment effects.[31]

[31] This framework can also be extended to test unconfoundedness assumptions, which motivate IPW-type attrition corrections

In the following, let S_i denote the stratum of individual i which has support $\mathcal{S}$, where $|\mathcal{S}| < \infty$.[32] To exclude trivial strata, we assume that $P(S_i = s) > 0$ for all $s \in \mathcal{S}$ throughout the paper. In a stratified randomized experiment, random assignment is defined by $(U_{i0}, U_{i1}, V_i) \perp T_i | S_i$, whereas in a completely randomized experiment this conditional independence assumption holds as an implication of simple randomization ($(S_i, U_{i0}, U_{i1}, V_i) \perp T_i$). As a result, the following proposition applies to both completely and stratified randomized experiments.

Proposition 6. *Assume* $(U_{i0}, U_{i1}, V_i) \perp T_i | S_i$.

(a) If $(U_{i0}, U_{i1}) \perp T_i | S_i, R_i$, *then*

(i) (Identification) $Y_{i1} | T_i = 0, S_i = s, R_i = 1 \stackrel{d}{=} Y_{i1}(0) | T_i = 1, S_i = s, R_i = 1$, *for* $s \in \mathcal{S}$.

(ii) (Sharp Testable Restriction) $Y_{i0} | T_i = 0, S_i = s, R_i = r \stackrel{d}{=} Y_{i0} | T_i = 1, S_i = s, R_i = r$ *for* $r = 0, 1$, $s \in \mathcal{S}$.

(b) If $(U_{i0}, U_{i1}) \perp R_i | T_i, S_i$, *then*

(i) (Identification) $Y_{i1} | T_i = \tau, S_i = s, R_i = 1 \stackrel{d}{=} Y_{i1}(\tau) | S_i = s$, *for* $\tau = 0, 1$, $s \in \mathcal{S}$.

(ii) (Sharp Testable Restriction) $Y_{i0} | T_i = \tau, S_i = s, R_i = r \stackrel{d}{=} Y_{i0}(0) | S_i = s$ *for* $\tau = 0, 1$, $r = 0, 1$, $s \in \mathcal{S}$.

The equality in (a.i) implies that we can identify the average treatment effect conditional on S for respondents as the difference in mean outcomes between treatment and control respondents in each stratum,

$$\begin{aligned} &E[Y_{i1}(1) - Y_{i1}(0) | T_i = 1, S_i = s, R_i = 1] \\ =&E[Y_{i1} | T_i = 1, S_i = s, R_i = 1] - E[Y_{i1} | T_i = 0, S_i = s, R_i = 1]. \text{ (CATE-R)} \end{aligned} \tag{3.3.8}$$

Alternatively, the ATE-R can then be identified by averaging over S_i, i.e.

$$\sum_{s \in \mathcal{S}} P(S_i = s | R_i = 1) \left(E[Y_{i1} | T_i = 1, S_i = s, R_i = 1] - E[Y_{i1} | T_i = 0, S_i = s, R_i = 1] \right)$$

The testable restriction in (a.ii) is the identity of the distribution of baseline outcome for treatment and control groups conditional on response status *and* stratum. In other words, the equality of the outcome distribution for treatment and control respondents (as well as for treatment and control attritors) conditional on stratrum is the sharp testable restriction of the IV-R assumption in the case of block randomization. The

(Huber, 2012), using baseline data. While interesting, this issue is outside the scope of the present paper.

[32] The finiteness of the number of strata motivates the finite-support assumption on $\mathcal{S}$. It is worth noting, however, that the results in the proposition hold for continuous conditioning variables as well.

results in part (b) of the proposition refer to IV-P in the context of block randomization. Thus, they are also conditional versions of the results in Proposition 5(b).

Randomization tests of the restrictions in Proposition 6(a.ii) and (b.ii) are provided in Section 3A, respectively. The key distinction between the randomization tests for stratified and completely randomized experiments is that in the former permutations are performed within strata.

3.3.3 Differential Attrition Rates and Internal Validity

The differential attrition rate test is the most widely used according to our review. Thus, we examine the relationship between internal validity and differential attrition rates ($P(R_i = 0|T_i = 1) \neq P(R_i = 0|T_i = 0)$). Our goal in this section is to formally understand the properties of the differential attrition rate test as a test of internal validity.

We first adapt the LATE framework (Imbens and Angrist, 1994; Angrist et al., 1996) to potential response. Specifically, in order to understand how treatment and control respondents and attritors consist of different response types, we modify the four types from the LATE literature: never-takers, always-takers, compliers and defiers. We establish four similar types as shown in Figure 3.2: never-responders ($(R_i(0), R_i(1)) = (0, 0)$), always-responders ($(R_i(0), R_i(1)) = (1, 1)$), treatment-only responders ($(R_i(0), R_i(1)) = (0, 1)$), and control-only responders ($(R_i(0), R_i(1)) = (1, 0)$).

Figure 3.2: Respondent and Attritor Subgroups

	Control ($T_i = 0$)	Treatment ($T_i = 1$)
Attritors ($R_i = 0$)	Treatment-only responders Never responders	Control-only responders Never responders
Respondents ($R_i = 1$)	Control-only responders Always responders	Treatment-only responders Always responders

We can now examine the attrition rates in the treatment and control group and how they relate to the different response types. By random assignment, the distribution of response types is identical across treatment and control groups, $(R_i(0), R_i(1)) \perp T_i$. In other words, the treatment and control groups consist of the same proportion of never responders, treatment-only responders, control-only responders and always responders, which we denote by p_{00}, p_{01}, p_{10} and p_{11}, respectively. With the aid of Figure 3.2, we note that the attrition rate in the control group equals the proportion of never-responders and treatment-only responders, whereas the attrition rate in the treatment group equals the proportion of never-responders and control-only responders, specifically

$$P(R_i = 0|T_i = 0) = p_{00} + p_{01}, \quad P(R_i = 0|T_i = 1) = p_{00} + p_{10}. \tag{3.3.9}$$

The difference in attrition rates across groups depends on the difference between the proportion of treatment-only and control-only responders, i.e. $P(R_i = 0|T_i = 0) - P(R_i = 0|T_i = 1) = p_{01} - p_{10}$. Thus, attrition rates are equal if the proportions of treatment-only and control-only responders are equal.

Next, we illustrate the relationship between differential attrition rates and the IV-R assumption (Proposition 5(a)), $(U_{i0}, U_{i1}) \perp T_i|R_i$. The proof of the proposition is given in Section 3C.

Proposition 7. *Suppose, in addition to* $(U_{i0}, U_{i1}, V_i) \perp T_i$*, one of the following is true,*

(i) $(U_{i0}, U_{i1}) \perp (R_i(0), R_i(1))$ *(Unobservables in* $Y \perp$ *Potential Response)*

(ii) $R_i(0) \leq R_i(1)$ *(wlog),* *(Monotonicity)*
& $P(R_i = 0|T_i) = P(R_i = 0)$ *(Equal Attrition Rates)*

(iii) $(U_{i0}, U_{i1})|R_i(0), R_i(1) \overset{d}{=} (U_{i0}, U_{i1})|R_i(0) + R_i(1)$ *(Exchangeability)*
& $P(R_i = 0|T_i) = P(R_i = 0)$ *(Equal Attrition Rates)*

then $(U_{i0}, U_{i1}) \perp T_i|R_i$.

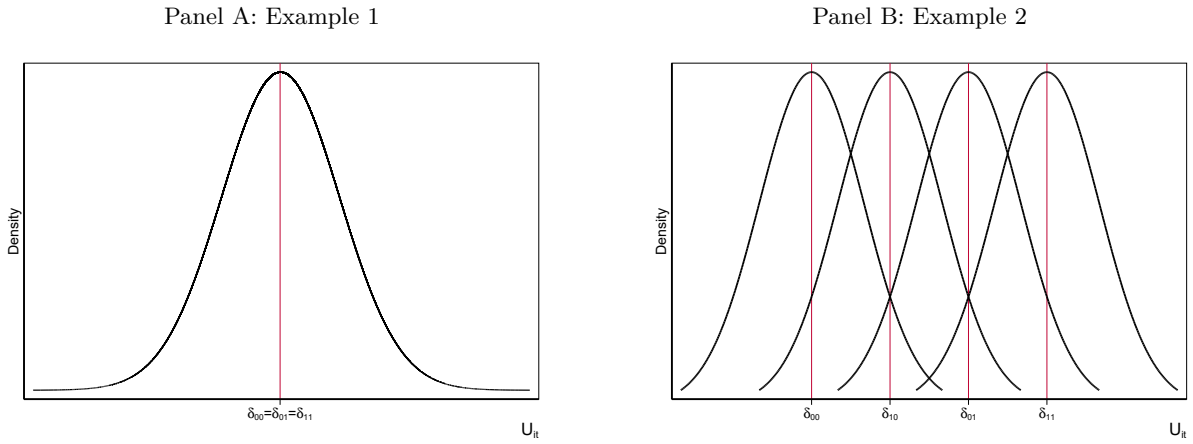
The main takeaway from the above proposition is that equal attrition rates alone do not constitute a sufficient condition for internal validity. Proposition 7(i) provides a case in which equal attrition rates are not necessary for internal validity. The assumption requires that all four treatment-response subgroups have the same unobservable distribution, which not only implies IV-R, but also IV-P, under random assignment. In the two other cases, (ii) and (iii), equal attrition rates together with an additional assumption imply the IV-R assumption. The monotonicity assumption in (ii) is from Lee (2009) and rules out control-only responders. The exchangeability restriction allows for both treatment-only and control-only responders, but it assumes that these two types have the same distribution of (U_{i0}, U_{i1}). This assumption may be plausible in experiments with two treatments.

Using these insights, we now provide two simple examples that illustrate that differential attrition rates can coincide with internal validity (*Example 1*) and that equal attrition rates can coincide with a violation of internal validity (*Example 2*). In Section 3.4, we design simulation experiments that mimic both examples to illustrate these points numerically. Furthermore, we find several empirical cases in Section 3.5 that are consistent with the theoretical conditions of *Example 1.*

Example 1. *(Internal Validity & Differential Attrition Rates)*
Assume that potential response satisfies monotonicity, i.e. $p_{10} = 0$*, and all response types have the same unobservable distribution,* $(U_{i0}, U_{i1}) \perp (R_i(0), R_i(1))$*. Panel A of Figure 3.3 illustrates the resulting distribution of* U_{it}*. By the above proposition, IV-P holds under random assignment, since* $(U_{i0}, U_{i1}) \perp (R_i(0), R_i(1)) \Rightarrow$

$(U_{i0}, U_{i1})|T_i, R_i \overset{d}{=} (U_{i0}, U_{i1})$. *Suppose that there is a group of individuals for whom it is too costly to respond if they are in the control group, so they only respond if assigned the treatment. Due to the presence of these treatment-only responders* ($p_{01} > 0$)*, the attrition rates in the treatment and control groups are not equal, specifically* $P(R_i = 0|T_i = 1) = p_{00}$*, and* $P(R_i = 0|T_i = 0) = p_{00} + p_{01}$*. This example thereby provides a case where we have differential attrition rates even though not only IV-R but also IV-P holds. Under these conditions, the differential attrition rate test would not control size as a test of internal validity as we illustrate in the simulation section.*

Figure 3.3: Distribution of U_{it} for Different Response Types



Notes: The above figure illustrates the distribution of U_{it} for the different subpopulations in Examples 1 and 2, where we assume $U_{it}|(R_i(0), R_i(1)) = (r_0, r_1) \overset{i.i.d.}{\sim} N(\delta_{r_0 r_1}, 1)$ for all $r_0, r_1 \in \{0, 1\}^2$ for $t = 0, 1$. Panel A represents Example 1 where we assume $(U_{i0}, U_{i1}) \perp (R_i(0), R_i(1))$, hence $\delta_{00} = \delta_{01} = \delta_{11}$. Panel B represents Example 2 where $\delta_{r_0 r_1}$ is unrestricted for $(r_0, r_1) \in \{0, 1\}^2$.

Example 2. *(Equal Attrition Rates & Violation of Internal Validity)*

Assume that potential response violates monotonicity, such that there are treatment-only and control-only responders,[33] *but their proportions are equal* ($p_{10} = p_{01} > 0$)*, which yields equal attrition rates across treatment and control groups.*[34] *If* $(U_{i0}, U_{i1}) \not\perp (R_i(0), R_i(1))$*, then the different response types will have different distributions of unobservables, as illustrated in Panel B of Figure 3.3. As a result, the distribution of* (U_{i0}, U_{i1}) *for treatment and control respondents defined in* (3C.2)-(3C.3) *will be different and hence IV-R*

[33] Violations of monotonicity are especially plausible in settings where we have two treatments. For the classical treatment-control case, a nice example of a violation of monotonicity of response is given in Glennerster and Takavarasha (2013). Suppose the treatment is a remedial program for public schools targeted toward students that have identified deficiencies in mathematics. Response in this setting is determined by whether students remain in the public school, which depends on their treatment status and initial mathematical ability, V_i. On one side, low-achieving students would drop out of school if they are assigned to the control group, but would remain in school if assigned the treatment. On the other side, parents of high-achieving students in the treatment group may be induced to switch their children to private schools because they are unhappy with the larger class sizes, while in the control group those students would remain in the public school. Furthermore, in the context of the LATE framework, de Chaisemartin (2017) provides several applications where monotonicity is implausible and establishes identification of a local average treatment effect under an alternative assumption.

[34] In the multiple treatment case, equal attrition rates are possible without requiring any two response types to have equal proportions in the population. See Section 3F in the online appendix for a derivation.

is violated.

A further limitation of the focus on the differential attrition rate test in empirical practice is that we cannot use it to test IV-P, even in cases where the differential attrition rate test is a valid test of IV-R. For instance, consider the case in which monotonicity holds and the attrition rates are equal across groups. We can then identify the ATE-R, since the respondent subpopulation is composed solely of always-responders as pointed out above. If the researchers are interested in identifying the treatment effect for the study population, however, they would have to test whether the always-responders are "representative" of the study population. To do so, one would have to test the restriction of the IV-P assumption in Proposition 5(b.ii).

3.3.4 Implications for Empirical Practice

Our theoretical analysis has multiple implications for empirical practice. For one, it underscores the importance of the object of interest in determining the appropriateness of an attrition test. Hence, explicitly stating the object of interest, whether it is the ATE-R, ATE, CATE-R or CATE, is important to justify a particular attrition test.

Our results further clarify the interpretation of attrition tests in the field experiment literature. The differential attrition rate test, which is implemented in 79% of papers in our review, is not based on a necessary condition of IV-R, and is not designed to test IV-P. Turning to the selective attrition tests, used in 60% of the papers, the null hypotheses are largely implications of the IV-R assumption (see Section 3E.2 in the online appendix). The most common version of this test (40% of all papers) uses respondents only; and hence, it does not exploit all the information in the baseline sample, specifically the attritors. Seventeen percent of papers do implement a selective attrition test that includes both respondents and attritors, suggesting that some authors are aware of the value of this information. Several of the null hypotheses they use, however, do not constitute IV-R or IV-P tests. This is perhaps unsurprising given the wide range of null hypotheses tested. Although authors do not in general conduct a direct test of IV-P, the inclusion of respondents and attritors in some selective attrition tests as well as the use of determinants of attrition tests suggest that some authors are likely interested in internally valid estimates for the study population. As we discuss in the empirical applications of Section 3.5, our results are promising for field experiments where treatment effects for the study population are of interest.

The Role of Covariates

An important question that arises in empirical practice is whether to include covariates in attrition tests. In our review of field experiments, we find that most authors use covariates in attrition tests regardless of the design of the study. While we restrict our review to experiments with baseline outcome data, there are settings where using covariates may be the only way to test attrition bias. In particular, some experiments target a population for which the baseline outcome always takes on the same value by design (e.g. if a job training program is targeted to unemployed people and employment is the main outcome). In other field experiments, baseline outcome data may not be available. We therefore provide a formal discussion of the role of covariates in attrition tests in this section.

Suppose that there is a set of covariates that are functions of the same determinants as the outcome, formally

$$W_{it} = \nu_t(U_{it}) \text{ for } t = 0, 1. \tag{3.3.10}$$

This definition pins down two types of covariates: (i) covariates that are themselves determinants of the outcome, i.e. $W_{it}^k = U_{it}^j$ for some k, j, $k = 1, \ldots, d_W$, $j = 1, \ldots, d_U$, or (ii) "proxy" variables, which are covariates determined by the same factors as the outcome Y_{it}. If this *a priori* information is true, the testable restrictions of the IV-R and IV-P assumptions would be on the joint distribution of $Z_{i0} = (Y_{i0}, W_{i0}')'$.[35] However, if this *a priori* information is false, then including covariates may lead to a false rejection of the identifying assumption in question. In addition, we note that studies that implement the selective attrition tests on all baseline variables, $\mathcal{Z}_{i0} = (Y_{i0}, W_{i0}', X_{i0}')'$, are testing the IV-R assumption for all variables in the survey as opposed to the outcome in question only. This IV-R assumption is a much stronger condition that may be violated, even if the IV-R assumption for the outcome in question holds.[36]

Thus, if the baseline survey contains determinants of the outcome or proxy variables (W_{i0}), then they can be included in tests of the IV-R or IV-P assumption for the outcome in question. Our results suggest, however, that the inclusion of covariates that are not determined solely by the same unobservables as the outcome (X_{i0}) may lead to false rejection of the IV-R or IV-P assumption. This outcome-specific approach to including other variables in attrition tests is further supported by our *Progresa* example, which illustrates empirically that attrition may affect internal validity differently for two different outcomes collected in the same survey. Another reason for potential over-rejection of internal validity in the literature is that a

[35] See Section 3B for details on regression tests for the multivariate case.

[36] Formally, the IV-R assumption relevant to all variables in the survey is $(\mathcal{E}_{i0}, \mathcal{E}_{i1}) \perp T_i | R_i$, where $\mathcal{Z}_{it} = \xi_t(\mathcal{E}_{it})$ and $\mathcal{E}_{it} = (U_{it}', \eta_{it}')'$. However, the IV-R assumption that ensures identification of treatment effects for the outcome in question is weaker, since it imposes the conditional random assignment restriction on the unobservables relevant to that outcome only, U_{it}.

substantial proportion of the implementation of selective attrition tests consists of individual tests for each baseline variable without correcting for multiple testing.

The implications of our analysis for empirical practice resonate with existing recommendations in the literature regarding the random assignment method used to ensure the similarity of treatment and control groups in terms of baseline observables in a given sample (i.e. "balance"). In seminal work on clinical trials, Altman (1985) emphasizes that imbalance should only concern the researcher if the variable in question relates to the outcome. Bruhn and McKenzie (2009) compare different stratified randomization procedures in terms of their ability to achieve balance. They point to the potential cost of using "irrelevant" variables in their simulation study and find that baseline outcome is by far the most informative determinant of future outcomes in various datasets.

Attrition Tests as Identification Tests

Our approach emphasizes that attrition tests are identification tests. While rejection of such tests is clear evidence against the identifying assumption in question, it is possible to fail to reject such tests when the assumption is in fact violated. This is because in general we can only test identifying assumptions by implication. In other words, their testable restrictions are necessary, but not sufficient for the identifying assumption to hold.[37] Figure 3.4 graphically presents this issue. The light gray area represents cases where the identifying assumption is violated yet the sharp testable restriction holds true.

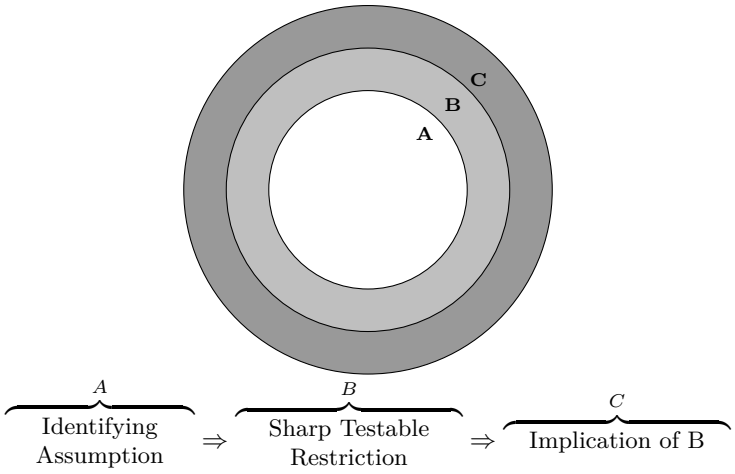
Figure 3.4 also illustrates that the sharp testable restriction is the strongest testable implication of the identifying assumption. Basing a test of the identifying assumption on another implication (C) leads to more cases where the implication holds yet the identifying assumption fails, represented by the dark gray area. Using sharp testable restrictions eliminates the cases in the dark gray area. The cases in the light gray area, which are unavoidable in general, complicate the interpretation of non-rejection of any identification test. Fortunately, our framework allows us to characterize the set of conditions under which this may or may not be a concern.

For both the IV-R and IV-P assumptions, there is a set of conditions in our setup under which identification holds if and only if the testable implication holds. These conditions consist of time homogeneity of the structural function and the unobservable distribution for the different treatment-response subpopulations (Chernozhukov et al., 2013).[38] This assumption may be plausible in some field experiments where researchers do not expect the structural function or the determinants of the outcome to vary between the baseline and follow-up surveys. To provide a simple example, suppose that the outcome equation is deter-

[37] In Footnote 28, we elaborate on why the theoretical case where the testable restriction is violated while identification holds is not empirically relevant in our setting.

[38] Formally, $\mu_0(d,u) = \mu_1(d,u)$ and $U_{i0}|T_i, R_i \overset{d}{=} U_{i1}|T_i, R_i$.

Figure 3.4: Graphical Illustration of Sharp Testable Restriction



mined by ability (U_i^1) and the opportunity cost of time (U_i^2), where the super-script is an index for the unobservables. We assume that both unobservables are time-invariant here to simplify notation. For a more general example with time-varying variables, see Section 3C.1. Now suppose that ability fulfils the IV-R assumption ($U_i^1 \perp T_i|R_i$), whereas the cost of time does not ($U_i^2 \not\perp T_i|R_i$). If ability *and* the cost of time both enter the baseline and follow-up outcomes, for instance,

$$Y_{i0} = U_i^1 + U_i^2$$
$$Y_{i1} = U_i^1 + U_i^2 + T_i(U_i^1 + U_i^2)$$

then comparisons between treatment and control respondents at follow-up would not be solely attributable to the treatment. Baseline outcome data would allow us to detect a violation of internal validity by comparing treatment and control respondents as well as treatment and control attritors.

Now let us consider a case where baseline outcome data would not help us detect such a violation of internal validity. This would require baseline outcome to only be a function of ability and not the cost of time, which only determines the outcome in the follow-up period,

$$Y_{i0} = U_i^1$$
$$Y_{i1} = U_i^1 + U_i^2 + T_i(U_i^1 + U_i^2).$$

Since ability fulfils the IV-R assumption, when comparing baseline outcome data of treatment and control respondents as well as treatment and control attritors, we would not detect any substantial differences

between these subgroups, even though internal validity is violated.[39] While we focus the example on the IV-R assumption, similar arguments can be made for the IV-P assumption.

A practical implication of our analysis is that when interpreting non-rejection of tests of the IV-R or IV-P assumptions, practitioners should consider whether the relationship between the outcome and its determinants may have changed over the time span between baseline and follow-up periods.

3.4 Simulation Study

We illustrate the theoretical results in the paper using a numerical study. The simulations examine the performance of the differential attrition rate test as well as both the mean and distributional tests of the IV-R and IV-P assumptions.

3.4.1 Simulation Design and Test Statistics

The data-generating process (DGP) is described in Panel A of Table 3.4. We assign individuals to one of the four response types: always-responders, never-responders, control-only responders, and treatment-only responders. The unobservables that determine the outcome consist of time-invariant and time-varying components. We introduce dependence between the unobservables in the outcome equation and potential response by allowing the means of the time-invariant component to differ for each response type. We also allow for heterogeneous treatment effects, so that the ATE-R can differ from the ATE.

We conduct simulations using four variants of this simulation design that feature different cases of IV-R and IV-P as summarized in Panel B of Table 3.4.[40] Designs I and II present cases where the differential rate test would have desirable properties as a test of IV-R.[41] Both designs allow for dependence between the unobservables in the outcome equation and potential response and impose monotonicity in the response equation by ruling out control-only responders. Design I allows for non-zero proportions of treatment-only responders and thereby a violation of IV-R. Design II rules out treatment-only responders and, as a result, we have IV-R, but not IV-P.

Designs III and IV illustrate *Examples 1* and *2* in Section 3.3.3, respectively. Design III demonstrates

[39] An interesting case that we illustrate in Section 3C.1 is that if the cost of time only interacts with the treatment, the difference in mean outcome between treatment and control respondents identifies an internally valid estimand that is not equal to the ATE-R.

[40] We only consider these four designs to keep the presentation clear. However, it is possible to combine different assumptions. For instance, if we assume $p_{01} = p_{10}$ and $(U_{i0}, U_{i1}) \perp (R_i(0), R_i(1))$, then we would have equal attrition rates and IV-P. We can also obtain a design that satisfies exchangeability by assuming $\delta_{01} = \delta_{10}$. If combined with $p_{01} = p_{10}$, then we would have equal attrition rates and IV-R only (Proposition 7(iii)).

[41] To be precise, in these designs, the differential attrition rate test would have non-trivial power when IV-R is violated while controlling size when IV-R holds.

Table 3.4: Simulation Design

Panel A. Data-Generating Process

Outcome:	$Y_{it} = \beta_1 D_{it} + \beta_2 D_{it}\alpha_i + \alpha_i + \eta_{it}$ for $t = 0, 1$ where $\beta_1 = \beta_2 = 0.25$.
Treatment:	$T_i \overset{i.i.d.}{\sim} Bernoulli(0.5)$, $D_{i0} = 0$, $D_{i1} = T_i$.
Response:	$R_i = (1 - T_i)R_i(0) + T_i R_i(1)$ where $p_{r_0 r_1} = P((R_i(0), R_i(1)) = (r_0, r_1))$ for $r_0, r_1 \in \{0, 1\}^2$
Unobservables:	$\begin{cases} U_{it} = (\alpha_i, \eta_{it})',\ t = 0, 1, \\ \alpha_i \vert R_i(0), R_i(1) \overset{i.i.d.}{\sim} \begin{cases} N(\delta_{00}, 1)\ if\ (R_i(0), R_i(1)) = (0, 0), \\ N(\delta_{01}, 1)\ if\ (R_i(0), R_i(1)) = (0, 1), \\ N(\delta_{10}, 1)\ if\ (R_i(0), R_i(1)) = (1, 0), \\ N(\delta_{11}, 1)\ if\ (R_i(0), R_i(1)) = (1, 1). \end{cases} \\ \eta_{i1} = 0.5\eta_{i0} + \epsilon_{i0},\ (\eta_{i0}, \epsilon_{i0})' \overset{i.i.d.}{\sim} N(0, 0.5 I_2) \end{cases}$

Panel B. Variants of the Design

Design	I	II	III	IV
Monotonicity in the Response Equation	Yes	Yes	Yes	No
Equal Attrition Rates	No	Yes	No	Yes
IV-R Assumption	No	Yes	Yes	No
IV-P Assumption ($(U_{i0}, U_{i1}) \perp R_i$)	No	No	Yes	No

Notes: For an integer k, I_k denotes a $k \times k$ identity matrix. In Designs I and II, we let $\delta_{00} = -0.5$, $\delta_{01} = 0.5$, and $\delta_{11} = -(\delta_{00}p_{00} + \delta_{01}p_{01})/p_{11}$, such that $E[\alpha_i] = 0$. In Design III, $\delta_{r_0 r_1} = 0$ for all $(r_0, r_1) \in \{0, 1\}^2$, which implies $U_{it} \perp (R_i(0), R_i(1))$ for $t = 0, 1$. In Design IV, $\delta_{00} = -0.5$, $\delta_{01} = -\delta_{10} = 0.25$, and $\delta_{11} = -(\delta_{00}p_{00} + \delta_{01}p_{01} + \delta_{10}p_{10})/p_{11}$. As for the proportions of the different subpopulations, in Designs I-III, we let $p_{00} = P(R_i = 0|T_i = 1)$, $p_{01} = P(R_i = 0|T_i = 0) - P(R_i = 0|T_i = 1)$, and $p_{11} = 1 - p_{00} - p_{01}$, whereas in Design IV, we fix $p_{10} = p_{01}$, $p_{00} = p_{10}/4$, and $P(R_i = 0|T_i = 0) = p_{00} + p_{10}$.

a setting in which we have differential attrition rates and IV-P. It imposes monotonicity and differential attrition rates as in Design I, but allows the unobservables in the outcome equation and potential response to be independent. Finally, Design IV follows *Example 2* in demonstrating a case in which there are equal attrition rates and a violation of internal validity. Here, we allow for a violation of monotonicity and dependence between the unobservables in the outcome equation and potential response. We impose that the proportion of treatment-only and control-only responders is identical and, as a result, the design features equal attrition rates.

In all four designs, we chose a range of attrition rates from the results of our review of the empirical literature (see Figure 3.1). Specifically, we allow for attrition rates in the control group from 5% to 30%, and differential attrition rates from zero to ten percentage points. To illustrate the implication of the designs for estimated mean effects, we report the simulation mean and standard deviation of the estimated difference in mean outcomes for the treatment and control respondents in the follow-up period ($\bar{Y}_1^{TR} - \bar{Y}_1^{CR}$).

The primary goal of our simulation analysis is to compare the performance of the differential attrition rate test as well as the mean and distributional IV-R and IV-P tests using a 5% level of significance. The differential attrition rate test is a two-sample t-test of the equality of attrition rates between the treatment and control group, $P(R_i = 0|T_i) = P(R_i = 0)$. The hypotheses of the mean IV-R and IV-P tests (denoted with an $\mathcal{M}$ subscript) are given by:

$$Y_{i0} = \gamma_{11} T_i R_i + \gamma_{01}(1 - T_i)R_i + \gamma_{10} T_i(1 - R_i) + \gamma_{00}(1 - T_i)(1 - R_i) + \epsilon_i \tag{3.4.1}$$

$$H_{0,\mathcal{M}}^{1,1}: \quad \gamma_{10} = \gamma_{00}, \qquad (CR\text{-}TR)$$

$$H_{0,\mathcal{M}}^{1,2}: \quad \gamma_{11} = \gamma_{01}, \qquad (CA\text{-}TA)$$

$$H_{0,\mathcal{M}}^{1}: \quad \gamma_{10} = \gamma_{00} \ \& \ \gamma_{11} = \gamma_{01}, \qquad (IV\text{-}R) \tag{3.4.2}$$

$$H_{0,\mathcal{M}}^{2}: \ \gamma_{11} = \gamma_{01} = \gamma_{10} = \gamma_{00}, \qquad (IV\text{-}P) \tag{3.4.3}$$

$H_{0,\mathcal{M}}^{1,1}$ ($H_{0,\mathcal{M}}^{1,2}$) tests the significance of mean differences between the treatment and control respondents (attritors) only. These two hypotheses are similar to widely used tests in the literature and are both implications of the IV-R assumption. $H_{0,\mathcal{M}}^{1}$ ($H_{0,\mathcal{M}}^{2}$) are the hypotheses of the mean IV-R (IV-P) tests in Section 3.3.2, which we implement using Wald statistics and asymptotic χ^2 critical values. To implement the distributional IV-R and IV-P tests, we use Kolmogorov-Smirnov-type (KS) statistics of their respective hypotheses,

$$H_0^1: \quad Y_{i0}|T_i, R_i = r \overset{d}{=} Y_{i0}|R_i = r, \text{ for } r = 0, 1, \tag{3.4.4}$$

$$H_0^2: \quad Y_{i0}|T_i, R_i \stackrel{d}{=} Y_{i0}. \tag{3.4.5}$$

We formally define the KS statistics for the above hypotheses in Section 3A.1, where we also describe the randomization procedures we use to obtain their p-values.

3.4.2 Simulation Results

Table 3.5 reports simulation rejection probabilities for the differential attrition rate test as well as the mean and distributional tests of the IV-R and IV-P assumptions for Designs I-IV. First, we consider the performance of the differential attrition rate test. Columns 1 through 3 of Table 3.5 report the simulation mean of the attrition rates for the control (C) and treatment (T) groups as well as the probability of rejecting a differential attrition rate test. Designs I and II, which obey monotonicity and allow for dependence between the unobservables in the outcome equation and potential response, illustrate the typical cases in which the differential attrition rate test can be viewed as a test of IV-R. In Design I, where internal validity is violated, the test rejects above 5%, while in Design II, where IV-R holds, the test controls size. Designs III and IV, on the other hand, illustrate the concerns we raise regarding the use of the differential attrition rate test as a test of IV-R. In Design III, the differential attrition rate test rejects at a frequency higher than 5% simply because the attrition rates are different even though IV-P holds. In Design IV, however, the differential attrition rate test does not reject above 5% when internal validity is violated because attrition rates are equal.

Next, we examine the performance of the IV-R tests, which are given in Columns 4 through 7 of Table 3.5. As expected, where IV-R holds (Designs II and III), the tests control size. Similarly, where IV-R is violated (Designs I and IV), the tests reject above 5%. In general, the relative power of the test statistics may differ depending on the DGP. In our simulation design, however, the rejection probabilities of the attritors-only test (CA-TA) and the joint tests ($Mean$ and KS) are significantly higher than the test based on the difference between the treatment and control respondents (CR-TR).[42]

The test statistics of the IV-P assumption (Columns 8 and 9 in Table 3.5) also behave according to our theoretical predictions. In Designs I, II and IV, where there is dependence between the unobservables in the outcome equation and potential response, the IV-P test rejects above 5%. Of particular interest is Design II, since internal validity holds for the respondents, but not for the population (i.e. IV-R holds, but IV-P does not). Thus, although the IV-P test does reject, the IV-R test does not reject above 5%. In this case, the difference in mean outcomes between treatment and control respondents (i.e. the estimated treatment

[42]This may be because the treatment-only responders are proportionately larger in the control attritor subgroup than in the treatment respondent subgroup.

effect) is not unbiased for the ATE (0.25), but it is internally valid for the respondents. In Design III, which is the only design where IV-P holds, both the mean and KS tests control size. Examining the difference in mean outcomes between treatment and control respondents at follow-up in this design, we find that it is unbiased for the ATE across all combinations of attrition rates.

Overall, the simulation results illustrate the limitations of the differential attrition rate test and show that the tests of the IV-R and IV-P assumptions we propose behave according to our theoretical predictions. For a more thorough numerical analysis of the finite-sample behavior of the Kolmogorov-Smirnov and Cramer-von-Mises statistics, see Section 3H in the online appendix.

3.5 Empirical Applications

To complement our simulation analysis, we apply the proposed tests of attrition bias to five published field experiments. The data comes from field experiments with both high attrition rates and publicly available data that includes attritors.[43] Thus, the exercise is not intended to draw inference about implications of applying various attrition tests to a representative sample of published field experiments. In addition, field experiments that are published in prestigious journals may not to be representative of all field experiment data, especially if perceptions of attrition bias had an impact on publication.

3.5.1 Implementation of Attrition Tests

Across the five selected articles included in this exercise, we conduct attrition tests for a total of 33 outcomes. This includes all outcomes with baseline data that are reported in the abstracts as well as all other unique outcomes with baseline data.[44] For each outcome included in this exercise, the appropriate attrition test depends on the type of outcome and the approach to randomization used in the experiment. For fully randomized experiments, we apply the tests of the IV-R and IV-P assumptions in Proposition 5. For stratified experiments, we instead apply the tests of the assumptions in Proposition 6.[45] For binary outcomes and also for all outcomes from clustered experiments, we apply regression-based mean tests (see Section 3B). For continuous outcomes in non-clustered experiments, we report p-values of the KS distributional tests using the appropriate randomization procedure.[46] For all tests, the results are presented in a way that is designed

[43] We selected the articles with the five highest attrition rates for which the data required to implement the attrition tests is available (see Section 3D.2 in the online appendix for details).

[44] If the article reports results separately by wave, we report attrition tests for each wave of a given outcome. We did not, however, report results for each heterogeneous treatment effect unless those results were reported in the abstract.

[45] When the number of strata in the experiment is larger than ten, we conduct a test with strata fixed effects only as opposed to the fully interacted regression in Section 3B in order to avoid high dimensional inference issues. Under the null, this specification is an implication of the sharp testable restrictions proposed in Proposition 6.

[46] We apply the Dufour (2006) randomization procedure to accommodate the possibility of ties.

to preserve the anonymity of the results and papers. Thus, attrition rates are presented as ranges, the results are not linked to specific articles, and we randomize the order of the outcomes such that they are not listed by paper.

In addition to applying our proposed attrition tests, we also consider how those tests might compare to other approaches. Thus, we apply a version of the tests commonly used in the literature to the data, including: the differential attrition rate test, the IV-R test using the respondent subsample only, and the IV-R test using the attritor subsample only. We use the same approaches to handling stratification and continuous outcomes in all three IV-R tests to ensure they are directly comparable, but that also means that we do not necessarily replicate the exact tests that are used in the articles from which we drew data for this exercise. Instead, we indicate whether authors' attrition tests reject for the outcomes for which they are available.

In keeping with our findings from Section 3.2, there is heterogeneity in the application of attrition tests across these articles. Two of the articles only report a differential attrition rate test, one article only reports a selective attrition test and two report both. The differential attrition rate test used by authors is based on survey-level attrition rates. As for the selective attrition test, each of the three articles that conducts such a test relies on a different implementation. One article uses a selective attrition test that neither constitutes an IV-R nor an IV-P test. The two other articles examine experiments that are randomized within strata. One article includes strata fixed effects in its selective attrition test in line with the IV-R tests implied by our analysis, whereas the other does not, and thus does not account for the stratification of the experimental design.

3.5.2 Results of the Empirical Applications

Our IV-R and IV-P test results reported in Table 3.6 have promising implications for the internal validity of randomized experiments. The joint IV-R test does not reject for any of the 33 outcomes at the 5% level. The IV-R tests using only respondents or attritors yield the same conclusion for all outcomes. Although there is often a substantial difference in the p-values for these two simple tests relative to the joint test for a given outcome, there is no consistent pattern in the direction of those differences. The IV-P test also does not reject the IV-P assumption at the 5% level for 26 out of the 33 outcomes (28 when accounting for multiple hypothesis testing).[47] While keeping in mind the usual caveats regarding the power of any test in finite samples, our results suggest that a researcher interested in treatment effects for the respondent subpopulation would not reject the relevant identifying assumption for any of the outcomes in our analysis,

[47]Although the number of outcomes from a given field experiment varies widely, the results are not driven by any one experiment or type of outcome.

even when exploiting all the information in the baseline sample (i.e. respondents and attritors). It is particularly notable that, for a majority of the outcomes we consider, a researcher would also not reject the identifying assumption that ensures the identification of the treatment effects for the study population.

Given its wide use in empirical practice, we also implement the differential attrition rate test. Using outcome-level attrition rates, we reject the null hypothesis of equal attrition rates at the 5% level for 9 of 33 outcomes (3 outcomes after correcting for multiple hypothesis testing).[48] For all 9 outcomes, the differential attrition rate test rejects the null hypothesis at the 5% level, whereas the IV-P assumption is not rejected at the 5% level using our test. These empirical cases are consistent with the testable implications of *Example 1*. Thus, according to our theoretical analysis, a researcher using the differential attrition rate test may falsely reject not only IV-R but also IV-P for these outcomes.

Next, we consider the results of the attrition tests reported by the authors (Table 3.6). The authors report a differential attrition rate test that is relevant to 30 out of the 33 outcomes and a selective attrition test for 8 outcomes. The reported differential attrition rate tests are rejected at the 5% level for 23 outcomes. The higher frequency of rejections of the authors' differential attrition rate test relative to ours is driven by their use of survey-level, as opposed to outcome-level, attrition rates. In the three articles in which the authors conduct a selective attrition test, they largely do not find evidence of selective attrition. They do, however, reject their version of the test at the 10% level for 2 of the 8 outcomes.

When we compare our test results with the authors', we note several differences. While we do not reject the IV-R assumption for any of the outcomes we consider, the authors reject their survey-level differential attrition rate test for 23 outcomes. Once we account for outcome-level attrition, we only reject equal attrition rates for 9 outcomes. As we note above, in all of these cases, our IV-P (or IV-R) test does not reject, which suggests that the differential attrition rate test is likely falsely rejecting internal validity for these outcomes. In addition, authors do not consistently account for the stratification of the experimental design in their selective attrition test, which may lead to a false rejection of internal validity.[49] Furthermore, one of the selective attrition tests used in the articles we examine does not constitute an IV-R or IV-P test. One limitation in comparing our results with the authors' is that, since they do not state their object of interest, it is not clear whether they intend to test for IV-R or IV-P.

Thus, we draw several conclusions from this empirical exercise. Our analysis illustrates that the differen-

[48] The relatively high differential attrition rates we find in this exercise are perhaps not surprising, given that overall attrition rates and differential attrition rates seem to be correlated, and these outcomes have fairly high attrition rates (McKenzie, 2019).

[49] To provide a simple example, consider a case where there are two strata (men and women). For simplicity, assume all men respond in the follow-up period. Now suppose 10% (5%) of women in the control (treatment) group do not respond to the follow-up survey, but the unobservables that affect outcome are independent of response. As a result, the treatment and control respondents consist of different proportions of men and women. It follows that, even though women in the different treatment-response subgroups have the same mean baseline outcome, the pooled treatment and control respondents may differ in that regard. Thus, a regression-based IV-R test that does not account for the stratification may falsely reject internal validity.

tial attrition rate test may lead to over-rejection of internal validity in practice. Furthermore, our empirical analysis highlights the disadvantages of the lack of consensus in empirical practice. Selective attrition tests are not universally implemented. The heterogeneity in the implementation of selective attrition tests could lead empirical researchers to unnecessarily question the internal validity of their study. In contrast, for all outcomes we consider, the results of our proposed joint IV-R test would not reject the identifying assumption that allows them to interpret their treatment effects as internally valid for the respondent subpopulation. If researchers were interested in the study population, our IV-P test results suggest that the data do not reject the identifying assumption in question for the majority of the outcomes in this exercise. Building on the *Progresa* example, our empirical exercise provides several additional cases where attrition impacts outcomes in the same experiment differently. These findings further highlight the advantages of our testing framework that allows the empirical researcher to align their attrition testing procedure with the outcome and population of interest.

3.6 Conclusion

This paper presents the problem of testing attrition bias in field experiments with baseline outcome data as an identification problem in a panel model. The proposed tests are based on the sharp testable restrictions of the identifying assumptions of the specific object of interest: either the average treatment effect for the respondents, the average treatment effect for the study population or a heterogeneous treatment effect. This study also provides theoretical conditions under which the differential attrition rate test, a widely used test, may not control size as a test of internal validity. The theoretical analysis has important implications for current empirical practice in testing attrition bias in field experiments. It also highlights that the majority of testing procedures used in the empirical literature have focused on the internal validity of treatment effects for the respondent subpopulation. The theoretical and empirical results, however, suggest that the treatment effects of the study population are important and possibly attainable in practice.

While this paper is a step forward toward understanding current empirical practice and establishing a standard in testing attrition bias in field experiments, it opens several questions for future research. Despite the availability of several approaches to correct for attrition bias (Lee, 2009; Huber, 2012; Behagel et al., 2015; Millán and Macours, 2019), alternative approaches that exploit the information in baseline outcome data as in the framework here may require weaker assumptions and hence constitute an important direction for future work. The extension of the analysis in this paper to the problem of attrition in the presence of partial compliance is another interesting direction. Furthermore, several practical aspects of the implementation of the proposed test may lead to pre-test bias issues. For instance, the proposed tests may be used in practice

to inform whether an attrition correction is warranted or not in the empirical analysis. Empirical researchers may also be interested in first testing the identifying assumption for treatment effects for the respondent subpopulation and then testing their validity for the entire study population. Inference procedures that correct for these and other pre-test bias issues are a priority for future work.

Finally, this paper has several policy implications. Attrition in a given study is often used as a metric to evaluate the study's reliability to inform policy. For instance, *What Works Clearinghouse*, an initiative of the U.S. Department of Education, has specific (differential) attrition rate standards for studies (IES, 2017). Our results indicate an alternative approach to assessing potential attrition bias. Furthermore, questions regarding external validity of treatment effects measured from field experiments are especially important from a policy perspective. This paper points to the possibility that in the presence of response problems, the identified effect in a given field experiment may only be valid for the respondent subpopulation, and hence may not identify the ATE for the study population. This is an important issue to consider when synthesizing results of field experiments to inform policy.

Table 3.5: Simulation Results on Differential Attrition Rates and Tests of Internal Validity ($ATE = 0.25$)

Design	Attrition Rates		Differential Attrition Rate Test	Tests of the IV-R Assumption				Tests of the IV-P Assumption		Difference in Mean Outcomes between Treatment & Control Respondents ($\bar{Y}_1^{TR} - \bar{Y}_1^{CR}$)		
				Mean Tests			KS Test	Mean Test	KS Test			
	C	T	$\hat{p}_{0.05}$	CR-TR	CA-TA	Joint	Joint	Joint	Joint	Mean	SD	$\hat{p}_{0.05}$
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Differential Attrition Rates + Monotonicity + $(U_{i0}, U_{i1}) \not\perp (R_i(0), R_i(1))$												
I	0.05	0.025	0.866	0.049	0.446	0.353	0.324	0.452	0.476	0.265	0.057	0.997
	0.10	0.05	0.995	0.076	0.719	0.635	0.582	0.792	0.787	0.282	0.058	0.998
	0.15	0.10	0.935	0.072	0.631	0.542	0.483	0.995	0.980	0.288	0.061	0.997
	0.20	0.15	0.867	0.072	0.532	0.442	0.412	1.000	1.000	0.296	0.063	0.996
	0.30	0.20	1.000	0.141	0.894	0.851	0.801	1.000	1.000	0.334	0.066	0.999
Equal Attrition Rates + Monotonicity + $(U_{i0}, U_{i1}) \not\perp (R_i(0), R_i(1))^\dagger$												
II	0.05	0.05	0.049	0.046	0.044	0.053	0.062	0.981	0.902	0.255	0.058	0.993
	0.10	0.10	0.053	0.043	0.045	0.045	0.056	1.000	0.999	0.262	0.060	0.991
	0.15	0.15	0.052	0.043	0.049	0.052	0.055	1.000	1.000	0.271	0.062	0.992
	0.20	0.20	0.049	0.045	0.047	0.050	0.050	1.000	1.000	0.280	0.064	0.990
	0.30	0.30	0.048	0.053	0.044	0.046	0.043	1.000	1.000	0.303	0.068	0.991
Differential Attrition Rates + Monotonicity + $(U_{i0}, U_{i1}) \perp (R_i(0), R_i(1))$ (*Example 1*)*												
III	0.05	0.025	0.866	0.055	0.051	0.056	0.052	0.065	0.050	0.248	0.058	0.990
	0.10	0.05	0.995	0.055	0.050	0.055	0.046	0.053	0.055	0.248	0.059	0.985
	0.15	0.10	0.935	0.057	0.052	0.053	0.045	0.053	0.059	0.247	0.061	0.983
	0.20	0.15	0.867	0.058	0.047	0.053	0.046	0.048	0.048	0.247	0.063	0.974
	0.30	0.20	1.000	0.057	0.053	0.052	0.043	0.049	0.048	0.248	0.066	0.964
Equal Attrition Rates + Violation of Monotonicity + $(U_{i0}, U_{i1}) \not\perp (R_i(0), R_i(1))$ (*Example 2*)												
IV	0.05	0.05	0.012	0.067	0.429	0.337	0.329	0.360	0.311	0.273	0.058	0.997
	0.10	0.10	0.013	0.131	0.708	0.653	0.577	0.708	0.582	0.302	0.059	0.999
	0.15	0.15	0.007	0.248	0.873	0.855	0.758	0.888	0.792	0.333	0.061	0.999
	0.20	0.20	0.004	0.422	0.934	0.951	0.859	0.970	0.913	0.367	0.063	0.999
	0.30	0.30	0.001	0.797	0.990	0.997	0.974	0.999	0.998	0.452	0.067	1.000

Notes: The above table reports simulation summary statistics for $n = 2{,}000$ across $2{,}000$ simulation replications. C denotes the control group, T denotes the treatment group, and $\hat{p}_{0.05}$ denotes the simulation rejection probability of a 5% test. The *Mean* tests of the IV-R (IV-P) assumption refer to the regression tests (Section 3B) of the null hypothesis in (3.4.2) ((3.4.3)). The KS statistics of the IV-R (IV- P) assumption are given in (3A.2) ((3A.4)), and their p-values are obtained using the proposed randomization procedures in Section 3A.1 ($B = 199$). The simulation mean, standard deviation (SD), and rejection probability of a two-sample t-test are reported for the difference in mean outcome between treatment and control respondents, $\bar{Y}_1^{TR} - \bar{Y}_1^{CR} = \frac{\sum_{i=1}^n Y_{i1} D_{i1} R_i}{\sum_{i=1}^n D_{i1} R_i} - \frac{\sum_{i=1}^n Y_{i1}(1-D_{i1}) R_i}{\sum_{i=1}^n (1-D_{i1}) R_i}$. All tests are conducted using $\alpha = 0.05$. Additional details of the design are provided in Table 3.4.

† (*) indicates IV-R only (IV-P).

Table 3.6: Attrition Tests Applied to Outcomes from Five Field Experiments

Outcome	Attrition Rate		Differential Attrition Rate Test	Tests of the IV-R Assumption			Test of the IV-P Assumption	Authors Reject the Null for:	
	Control (%)	Differential (percentage points)		CR-TR	CA-TA	Joint	Joint	Differential Attrition Rates Test	Selective Attrition Test
1	[10 - 30]	(10 - 20]	0.025†	0.567	0.948	0.832	0.563	Yes: 5%	No
2	[10 - 30]	(0 - 5]	0.887	0.514	0.546	0.571	0.60	No	Yes: 10%
3	[10 - 30]	(0 - 5]	0.109	0.834	0.751	0.879	0.956	Yes: 5%	–
4	[10 - 30]	(0 - 5]	0.486	0.351	0.701	0.576	0.000*	Yes: 5%	–
5	[10 - 30]	(0 - 5]	0.100	0.421	0.526	0.668	0.755	Yes: 5%	–
6	[10 - 30]	(0 - 5]	0.086	0.392	0.098	0.187	0.313	Yes: 5%	–
7	[10 - 30]	(0 - 5]	0.056	0.315	0.575	0.490	0.652	Yes: 5%	–
8	[10 - 30]	(0 - 5]	0.027	0.359	0.381	0.537	0.679	Yes: 5%	–
9	[10 - 30]	(0 - 5]	0.129	0.190	0.532	0.312	0.008*	Yes: 5%	–
10	[30 - 50]	(0 - 5]	0.301	0.202	0.191	0.198	0.002*	Yes: 5%	–
11	[10 - 30]	(0 - 5]	0.030	0.688	0.966	0.917	0.979	Yes: 5%	–
12	[10 - 30]	(0 - 5]	0.955	0.120	0.114	0.250	0.000*	No	–
13	[10 - 30]	(10 - 20]	0.039†	0.827	0.120	0.277	0.441	Yes: 5%	–
14	[10 - 30]	(0 - 5]	0.788	0.861	0.194	0.423	0.525	No	–
15	[10 - 30]	(10 - 20]	0.048†	0.682	0.558	0.800	0.609	Yes: 5%	No
16	[10 - 30]	(0 - 5]	0.798	0.802	0.180	0.404	0.590	No	No
17	[10 - 30]	(10 - 20]	0.037†	0.685	0.428	0.711	0.843	Yes: 5%	–
18	[10 - 30]	(0 - 5]	0.784	0.833	0.169	0.384	0.546	No	–
19	[30 - 50]	(0 - 5]	0.127	0.700	0.494	0.690	0.010*	Yes: 5%	–
20	[30 - 50]	(0 - 5]	0.241	0.605	0.476	0.720	0.697	Yes: 5%	–
21	[10 - 30]	(0 - 5]	0.084	0.796	0.261	0.518	0.671	Yes: 5%	–
22	[30 - 50]	(0 - 5]	0.218	0.748	0.183	0.385	0.022†	Yes: 5%	–
23	[30 - 50]	(0 - 5]	0.128	0.328	0.632	0.615	0.053	Yes: 5%	–
24	[30 - 50]	(0 - 5]	0.134	0.133	0.976	0.337	0.528	Yes: 5%	–
25	[30 - 50]	(0 - 5]	0.118	0.718	0.510	0.707	0.029†	Yes: 5%	–
26	[30 - 50]	(0 - 5]	0.348	0.663	0.370	0.691	0.807	Yes: 5%	–
27	[30 - 50]	(0 - 5]	0.217	0.883	0.768	0.858	0.423	Yes: 5%	–
28	[10 - 30]	(0 - 5]	0.061	0.218	0.986	0.518	0.609	Yes: 5%	–
29	[10 - 30]	(5 - 10]	0.036*	0.276	0.698	0.832	0.106	–	–
30	[10 - 30]	(10 - 20]	0.000*	0.354	0.984	0.864	0.064	–	No
31	[30 - 50]	(10 - 20]	0.047*	0.144	0.440	0.526	0.692	–	Yes: 10%
32	[10 - 30]	(0 - 5]	0.867	0.580	0.509	0.798	0.720	No	No
33	[10 - 30]	(5 - 10]	0.437	0.421	0.887	0.683	0.447	No	No

Notes: The table reports p-values for the differential attrition rate test as well as tests of the IV-R and IV-P assumptions. The symbol * (†) next to the p-value indicates that the relevant test statistic remains statistically significant after applying the Benjamini-Hochberg correction at 5% (10%) for outcomes from the same article (see Benjamini and Hochberg (1995) for details on this procedure). $CR - TR$ ($CA - TA$) indicates difference across treatment and control respondents (attritors). Joint tests include all four treatment-response sub-groups. Regression tests are implemented for (i) the differential attrition rate test, (ii) for the IV-R and IV-P tests with binary outcomes, and (iii) for cluster-randomized trials. Standard errors are clustered (if treatment is randomized at the cluster level) and strata fixed effects are included (if treatment is randomized within strata). For continuous outcomes in non-clustered trials, p-values of the KS tests are implemented using the appropriate randomization procedures ($B = 499$). For stratified experiments with less than ten strata, the test proposed in Proposition 6 is implemented. The last two columns of the table report whether (and the significance level at which) the authors reject their tests of differential attrition rates and selective attrition, respectively. The dash indicates that the test was not reported by the authors.

3.7 References

A. Abadie, M. M. Chingos, and M. R. West. Endogenous stratification in randomized experiments. *Review of Economics and Statistics*, 100(4):567–580, 2018. doi: 10.1162/rest_a_00732. URL https://doi.org/10.1162/rest_a_00732.

H. Ahn and J. L. Powell. Semiparametric estimation of censored selection models with a nonparametric selection mechanism. *Journal of Econometrics*, 58(1):3–29, 1993. ISSN 0304-4076. doi: https://doi.org/10.1016/0304-4076(93)90111-H. URL http://www.sciencedirect.com/science/article/pii/030440769390111H.

D. G. Altman. Comparability of randomised groups. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 34(1):125–136, 1985. doi: 10.2307/2987510. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.2307/2987510.

J. Altonji and R. Matzkin. Cross-section and panel data estimators for nonseparable models with endogenous regressors. *Econometrica*, 73(3):1053–1102, 2005.

I. Andrews and E. Oster. A simple approximation for evaluating external validity bias. *Economics Letters*, 178:58 – 62, 2019. ISSN 0165-1765. doi: https://doi.org/10.1016/j.econlet.2019.02.020. URL http://www.sciencedirect.com/science/article/pii/S0165176519300655.

J. D. Angrist, G. W. Imbens, and D. B. Rubin. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434):444–455, 1996. doi: 10.1080/01621459.1996.10476902. URL http://www.tandfonline.com/doi/abs/10.1080/01621459.1996.10476902.

S. Athey and G. Imbens. Chapter 3 - the econometrics of randomized experiments. In A. V. Banerjee and E. Duflo, editors, *Handbook of Field Experiments*, volume 1 of *Handbook of Economic Field Experiments*, pages 73 – 140. North-Holland, 2017. doi: https://doi.org/10.1016/bs.hefe.2016.10.003. URL http://www.sciencedirect.com/science/article/pii/S2214658X16300174.

S. Athey, D. Eckles, and G. W. Imbens. Exact p-values for network interference. *Journal of the American Statistical Association*, 113(521):230–240, 2018. doi: 10.1080/01621459.2016.1241178. URL https://doi.org/10.1080/01621459.2016.1241178.

T. Azzam, M. Bates, and D. Fairris. Do learning communities increase first year college retention? testing the external validity of randomized control trials. Unpublished, 2018.

S. Baird, J. A. Bohren, C. McIntosh, and B. Özler. Optimal design of experiments in the presence of interference. *Review of Economics and Statistics*, 100(5):844–860, 2018. doi: 10.1162/rest_a_00716. URL https://doi.org/10.1162/rest_a_00716.

G. Barrett, P. Levell, and K. Milligan. A Comparison of Micro and Macro Expenditure Measures across Countries Using Differing Survey Methods. In *Improving the Measurement of Consumer Expenditures*, NBER Chapters, pages 263–286. National Bureau of Economic Research, Inc, 2014. URL https://ideas.repec.org/h/nbr/nberch/12665.html.

L. Behagel, B. Crépon, M. Gurgand, and T. L. Barbanchon. Please call again: Correcting nonresponse bias in treatment effect models. *Review of Economics and Statistics*, 97:1070–1080, 2015.

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995. ISSN 00359246. URL http://www.jstor.org/stable/2346101.

C. A. Bester and C. Hansen. Identification of marginal effects in a nonparametric correlated random effects model. *Journal of Business and Economic Statistics*, 27(2):235–250, 2009.

M. Bruhn and D. McKenzie. In pursuit of balance: Randomization in practice in development field experiments. *American Economic Journal: Applied Economics*, 1(4):200–232, October 2009. doi: 10.1257/app.1.4.200. URL http://www.aeaweb.org/articles?id=10.1257/app.1.4.200.

F. A. Bugni, I. A. Canay, and A. M. Shaikh. Inference under covariate-adaptive randomization. *Journal of the American Statistical Association*, 113(524):1784–1796, 2018. doi: 10.1080/01621459.2017.1375934. URL https://doi.org/10.1080/01621459.2017.1375934.

I. A. Canay, J. P. Romano, and A. M. Shaikh. Randomization tests under an approximate symmetry assumption. *Econometrica*, 85(3):1013–1030, 2017. doi: 10.3982/ECTA13081. URL https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA13081.

X. Chen and C. A. Flores. Bounds on treatment effects in the presence of sample selection and noncompliance: The wage effects of job corps. *Journal of Business & Economic Statistics*, 33(4):523–540, 2015. doi: 10.1080/07350015.2014.975229. URL https://doi.org/10.1080/07350015.2014.975229.

V. Chernozhukov, I. Fernandez-Val, J. Hahn, and W. Newey. Average and quantile effects in nonseparable panel data models. *Econometrica*, 81(2):pp.535–580, 2013.

M. Das, W. K. Newey, and F. Vella. Nonparametric estimation of sample selection models. *Review of Economic Studies*, 70(1):33–58, 2003. doi: 10.1111/1467-937X.00236. URL +http://dx.doi.org/10.1111/1467-937X.00236.

C. de Chaisemartin. Tolerating defiance? local average treatment effects without monotonicity. *Quantitative Economics*, 8(2):367–396, 2017. doi: 10.3982/QE601. URL https://onlinelibrary.wiley.com/doi/abs/10.3982/QE601.

C. de Chaisemartin and L. Behaghel. Estimating the Effect of Treatments Allocated by Randomized Waiting Lists. Papers 1511.01453, arXiv.org, Nov. 2018. URL https://ideas.repec.org/p/arx/papers/1511.01453.html.

J.-M. Dufour. Monte carlo tests with nuisance parameters: A general approach to finite-sample inference and nonstandard asymptotics. *Journal of Econometrics*, 133(2):443 – 477, 2006. ISSN 0304-4076. doi: https://doi.org/10.1016/j.jeconom.2005.06.007. URL http://www.sciencedirect.com/science/article/pii/S0304407605001260.

J.-M. Dufour, A. Farhat, L. Gardiol, and L. Khalaf. Simulation-based finite sample normality tests in linear regressions. *Econometrics Journal*, 1(1):154–173, 1998. doi: 10.1111/1368-423X.11009. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/1368-423X.11009.

H. Fricke, M. Fröhlich, M. Huber, and M. Lechner. Endogeneity and non-response bias in treatment evaluation: Nonparametric identification of causal effects by instruments. IZA Discussion Papers, No. 9428, Institute for the Study of Labor (IZA), Bonn, 2015.

D. Ghanem. Testing identifying assumptions in nonseparable panel data models. *Journal of Econometrics*, 197:202–217, 2017.

R. Glennerster and K. Takavarasha. *Running Randomized Evaluations: A Practical Guide*. Princeton University Press, student edition edition, 2013. ISBN 9780691159270. URL http://www.jstor.org/stable/j.ctt4cgd52.

J. A. Hausman and D. A. Wise. Attrition bias in experimental and panel data: The gary income maintenance experiment. *Econometrica*, 47(2):455–473, 1979. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/1914193.

J. J. Heckman. The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models. In Sanford V. Berg, editor, *Annals of Economic and Social Measurement*, volume 5, pages 475–492. National Bureau of Economic Research, 1976. URL https://www.nber.org/chapters/c10491.pdf.

J. J. Heckman. Sample selection bias as a specification error. *Econometrica*, 47(1):153–161, 1979. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/1912352.

K. Hirano, G. W. Imbens, G. Ridder, and D. B. Rubin. Combining panel data sets with attrition and refreshment samples. *Econometrica*, 69(6):1645–1659, 2001. doi: 10.1111/1468-0262.00260. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/1468-0262.00260.

K. Hirano, G. W. Imbens, and G. Ridder. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189, 2003. doi: 10.1111/1468-0262.00442. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/1468-0262.00442.

S. Hoderlein and H. White. Nonparametric identification of nonseparable panel data models with generalized fixed effects. *Journal of Econometrics*, 168(2):300–314, 2012.

J. L. Horowitz and C. F. Manski. Nonparametric analysis of randomized experiments with missing covariate and outcome data. *Journal of the American Statistical Association*, 95(449):77–84, 2000. doi: 10.1080/01621459.2000.10473902. URL https://amstat.tandfonline.com/doi/abs/10.1080/01621459.2000.10473902.

D. G. Horvitz and D. J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952. doi: 10.1080/01621459.1952.10483446. URL https://amstat.tandfonline.com/doi/abs/10.1080/01621459.1952.10483446.

Y.-C. Hsu, C.-A. Liu, and X. Shi. Testing generalized regression monotonicity. *Econometric Theory*, page 1 – 55, 2019. doi: 10.1017/S0266466618000439.

M. Huber. Identification of average treatment effects in social experiments under alternative forms of attrition. *Journal of Educational and Behavioral Statistics*, 37(3):443–474, 2012. doi: 10.3102/1076998611411917. URL https://doi.org/10.3102/1076998611411917.

IES. What Works Clearinghouse. Standards Handbook Version 4.0. Technical report, U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, What Works Clearinghouse, 2017.

G. W. Imbens and J. D. Angrist. Identification and estimation of local average treatment effects. *Econometrica*, 62(2):467–475, 1994. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/2951620.

G. W. Imbens and D. B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, 2015. doi: 10.1017/CBO9781139025751.

INSP. General Rural Methodology Note. Technical report, Instituto Nacional de Salud Publica, 2005. URL https://evaluacion.prospera.gob.mx/es/wersd53465sdg1/eval{_}cuant/general{_}rural{_}methodology{_}note{_}2005.pdf.

M. Kasy and A. Sautmann. Adaptive treatment assignment in experiments for policy choice. *Econometrica*, Forthcoming, 2020.

T. Kitagawa. A test for instrument validity. *Econometrica*, 83(5):2043–2063, 2015. doi: 10.3982/ECTA11974. URL https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA11974.

P. Kline and A. Santos. Sensitivity to missing data assumptions: Theory and an evaluation of the u.s. wage structure. *Quantitative Economics*, 4(2):231–267, 2013. doi: 10.3982/QE176. URL https://onlinelibrary.wiley.com/doi/abs/10.3982/QE176.

D. S. Lee. Training, wages, and sample selection: Estimating sharp bounds on treatment effects. *Review of Economic Studies*, 76(3):1071–1102, 2009. doi: 10.1111/j.1467-937X.2009.00536.x. URL +http://dx.doi.org/10.1111/j.1467-937X.2009.00536.x.

E. L. Lehmann and J. P. Romano. *Testing Statistical Hypotheses*. Springer, New York, third edition, 2005. ISBN 0-387-98864-5.

C. F. Manski. Partial identification with missing data: Concepts and findings. *International Journal of Approximate Reasoning*, 39(2):151 – 165, 2005. ISSN 0888-613X. doi: https://doi.org/10.1016/j.ijar.2004.10.006. URL http://www.sciencedirect.com/science/article/pii/S0888613X04001124.

D. McKenzie. Beyond baseline and follow-up: The case for more t in experiments. *Journal of Development Economics*, 99(2):210–221, 2012. ISSN 0304-3878. doi: https://doi.org/10.1016/j.jdeveco.2012.01.002. URL http://www.sciencedirect.com/science/article/pii/S030438781200003X.

D. McKenzie. Attrition rates typically aren't that different for the control group than the treatment group – really? and why? *Development Impact Blog*, January 07, 2019. https://blogs.worldbank.org/impactevaluations/attrition-rates-typically-aren-t-different-control-group-treatment-group-really-and-why.

B. D. Meyer, W. K. C. Mok, and J. X. Sullivan. Household surveys in crisis. *Journal of Economic Perspectives*, 29(4):199–226, November 2015. doi: 10.1257/jep.29.4.199. URL http://www.aeaweb.org/articles?id=10.1257/jep.29.4.199.

T. M. Millán and K. Macours. Attrition in randomized control trials: Using tracking information to correct bias. Unpublished Manuscript, 2019.

I. Mourifié and Y. Wan. Testing local average treatment effect assumptions. *Review of Economics and Statistics*, 99(2):305–313, 2017.

K. Muralidharan, M. Romero, and K. Wüthrich. Factorial designs, model selection, and (incorrect) inference in randomized experiments. Working Paper 26562, National Bureau of Economic Research, December 2019. URL http://www.nber.org/papers/w26562.

J. M. Robins, A. Rotnitzky, and L. P. Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866, 1994.

D. B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976. ISSN 00063444. URL http://www.jstor.org/stable/2335739.

E. Skoufias. Progresa and its impacts on the welfare of rural households in mexico. Research Report 139, International Food Policy Research Institute (IFPRI), 2005. URL https://ideas.repec.org/p/fpr/resrep/139.html.

G. Vazquez-Bare. Identification and estimation of spillover effects in randomized experiments. Unpublished Manuscript, 2020.

J. M. Wooldridge. Selection corrections for panel data models under conditional mean independence assumptions. *Journal of Econometrics*, 68(1):115 – 132, 1995. ISSN 0304-4076. doi: https://doi.org/10.1016/0304-4076(94)01645-G. URL http://www.sciencedirect.com/science/article/pii/030440769401645G.

A. Young. Channeling Fisher: Randomization Tests and the Statistical Insignificance of Seemingly Significant Experimental Results*. *Quarterly Journal of Economics*, 134(2):557–598, 11 2018. ISSN 0033-5533. doi: 10.1093/qje/qjy029. URL https://doi.org/10.1093/qje/qjy029.

3A Randomization Tests of Internal Validity

We present randomization procedures to test the IV-R and IV-P assumptions for completely and stratified randomized experiments. The proposed procedures approximate the exact p-values of the proposed distributional statistics under the cross-sectional i.i.d. assumption when the outcome distribution is continuous.[50] They can also be adapted to accommodate possibly discrete or mixed outcome distributions, which may result from rounding or censoring in the data collection, by applying the procedure in Dufour (2006). In this section, we focus on distributional statistics for the testable restrictions on the baseline outcome as in Propositions 5 and 6. The randomization procedures we propose, however, can be applied to test joint distributional hypotheses that include covariates as in Section 3.3.4.

We first outline a general randomization procedure that we adapt to the different settings we consider.[51] Given a dataset $\mathbf{Z}$ and a statistic $T_n = T(\mathbf{Z})$ that tests a null hypothesis H_0, we use the following procedure to provide a stochastic approximation of the exact p-value for the test statistic T_n exploiting invariant transformations $g \in \mathcal{G}_0$ (Lehmann and Romano, 2005, Chapter 15.2). Specifically, the transformations $g \in \mathcal{G}_0$ satisfy $\mathbf{Z} \overset{d}{=} g(\mathbf{Z})$ under H_0 only.

Procedure 1. *(Randomization)*

1. *For g_b, which is i.i.d. Uniform($\mathcal{G}_0$), compute $\hat{T}_n(g_b) = T(g_b(\boldsymbol{Z}))$,*
2. *Repeat Step 1 for $b = 1, \ldots, B$ times,*
3. *Compute the p-value, $\hat{p}_{n,B} = \frac{1}{B+1}\left(1 + \sum_{b=1}^{B} 1\{\hat{T}_n(g_b) \geq T_n\}\right)$.*

A test that rejects when $\hat{p}_{n,B} \leq \alpha$ is level α for any B (Lehmann and Romano, 2005, Chapter 15.2). In our application, the invariant transformations in $\mathcal{G}_0$ consist of permutations of individuals across certain subgroups in our data set. The subgroups are defined by the combination of response and treatment in the case of completely randomized trials, and all the combinations of response, treatment, and stratum in the case of trials that are randomized within strata.

3A.1 Completely Randomized Trials

The testable restriction of the IV-R assumption, stated in Proposition 5(a.ii), implies that the distribution of baseline outcome is identical for treatment and control respondents as well as treatment and control attritors.

[50] We maintain the cross-sectional i.i.d. assumption to simplify the presentation. The randomization procedures proposed here remain valid under weaker exchangeability-type assumptions.

[51] See Lehmann and Romano (2005); Canay et al. (2017) for a more detailed review.

Thus, the joint hypothesis is given by

$$H_0^1 : F_{Y_{i0}|T_i=0,R_i=r} = F_{Y_{i0}|T_i=1,R_i=r} \text{ for } r = 0, 1. \tag{3A.1}$$

The general form of the distributional statistic for *each* of the equalities in the null hypothesis above is

$$T_{n,r}^1 = \left\| \sqrt{n} \left(F_{n,Y_{i0}|T_i=0,R_i=r} - F_{n,Y_{i0}|T_i=1,R_i=r} \right) \right\| \quad \text{for } r = 0, 1,$$

where for a random variable X_i, F_{n,X_i} denotes the empirical cdf, i.e. the sample analogue of F_{X_i}, and $\|.\|$ denotes some non-random or random norm. Different choices of the norm give rise to different statistics. For instance, the KS and CM statistics are the most widely known and used. The former is obtained by using the L^∞ norm over the sample points, i.e. $\|f\|_{n,\infty} = \max_i |f(y_i)|$, whereas the latter is obtained by using an L^2 norm, i.e. $\|f\|_{n,2} = \sum_{i=1}^n f(y_i)^2/n$. In order to test the *joint* hypothesis in (3A.1), the two following statistics that aggregate over $T_{n,r}^1$ for $r = 0, 1$ are standard choices in the literature (Imbens and Rubin, 2015),[52]

$$\begin{aligned} T_{n,m}^1 &= \max\{T_{n,0}^1, T_{n,1}^1\}, \\ T_{n,p}^1 &= p_{n,0} T_{n,0}^1 + p_{n,1} T_{n,1}^1, \quad \text{where } p_{n,r} = \sum_{i=1}^n 1\{R_i = r\}/n \text{ for } r = 0, 1. \end{aligned}$$

The joint KS statistic we use to test H_0^1 in the simulation and empirical section is given by

$$\begin{aligned} KS_{n,m}^1 &= \max\{KS_{n,0}^1, KS_{n,1}^1\}, \text{where for } r = 0, 1 \\ KS_{n,r}^1 &= \max_{i:R_i=r} \left| \sqrt{n} \left(F_{n,Y_{i0}}(y_{i0}|T_i = 1, R_i = r) - F_{n,Y_{i0}}(y_{i0}|T_i = 0, R_i = r) \right) \right|. \end{aligned} \tag{3A.2}$$

Let $\mathcal{G}_0^1$ denote the set of all permutations of individual observations within respondent and attritor subgroups, for $g \in \mathcal{G}_0^1$, $g(\mathbf{Z}) = \{(Y_{i0}, T_{g(i)}, R_{g(i)}) : R_{g(i)} = R_i, 1 \leq i \leq n\}$. Under H_0^1 and the cross-sectional i.i.d. assumption, $\mathbf{Z} \stackrel{d}{=} g(\mathbf{Z})$ for $g \in \mathcal{G}_0^1$. Hence, we can obtain p-values for $T_{n,m}^1$ and $T_{n,p}^1$ under H_0^1 by applying Procedure 1 using the set of permutations $\mathcal{G}_0^1$.

We now consider testing the restriction of the IV-P assumption stated in Proposition 5(b.ii). This restriction implies that the distribution of the baseline outcome variable is identically distributed across all four subgroups defined by treatment and response status. Let $(T_i, R_i) = (\tau, r)$, where $(\tau, r) \in \mathcal{T} \times \mathcal{R} = \{(0,0), (0,1), (1,0), (1,1)\}$ and (τ_j, r_j) denote the j^{th} element of $\mathcal{T} \times \mathcal{R}$. Then, the joint hypothesis is given

[52]There are other possible approaches to construct joint statistics. We compare the finite-sample performance of the two joint statistics we consider numerically in Section 3H of the online appendix.

wlog by

$$H_0^2 : F_{Y_{i0}|T_i=\tau_j,R_i=r_j} = F_{Y_{i0}|T_i=\tau_{j+1},R_i=r_{j+1}} \text{ for } j = 1,\ldots,|\mathcal{T}\times\mathcal{R}|-1. \tag{3A.3}$$

In this case, the two statistics that we propose to test the *joint* hypothesis are:

$$\begin{aligned} T_{n,m}^2 &= \max_{j=1,\ldots,|\mathcal{T}\times\mathcal{R}|-1} \left\| \sqrt{n}\left(F_{n,Y_{i0}|T_i=\tau_j,R_i=r_j} - F_{n,Y_{i0}|T_i=\tau_{j+1},R_i=r_{j+1}}\right)\right\|, \\ T_{n,p}^2 &= \sum_{j=1}^{|\mathcal{T}\times\mathcal{R}|-1} w_j \left\| \sqrt{n}\left(F_{n,Y_{i0}|T_i=\tau_j,R_i=r_j} - F_{n,Y_{i0}|T_i=\tau_{j+1},R_i=r_{j+1}}\right)\right\| \end{aligned}$$

for some fixed or data-dependent non-negative weights w_j for $j = 1,\ldots,|\mathcal{T}\times\mathcal{R}|-1$. In the simulation and empirical sections, we use the following KS statistic to test H_0^2

$$\begin{aligned} KS_n^2 &= \max_{j=1,2,3} KS_{n,j}^2, \text{ where} \\ KS_{n,j}^2 &= \max_i \left|\sqrt{n}\left(F_{n,Y_{i0}}(y_{i0}|T_i=\tau_j,R_i=r_j) - F_{n,Y_{i0}}(y_{i0}|T_i=\tau_{j+1},R_i=r_{j+1})\right)\right|. \end{aligned} \tag{3A.4}$$

and $\{\tau_j, r_j\}$ is the j^{th} element of $\mathcal{T}\times\mathcal{R} = \{(0,0),(0,1),(1,0),(1,1)\}$.

Under H_0^2 and the cross-sectional i.i.d. assumption, any random permutation of individuals across the four treatment-response subgroups will yield the same joint distribution of the data. Specifically, for $g \in \mathcal{G}_0^2$, $g(\mathbf{Z}) = \{(Y_{i0}, T_{g(i)}, R_{g(i)}) : 1 \le i \le n\}$. We can hence apply Procedure 1 using $\mathcal{G}_0^2$ to obtain approximately exact p-values for the statistic $T_{n,m}^2$ or $T_{n,p}^2$ under H_0^2.

3A.2 Stratified Randomized Trials

As pointed out in Section 3.3.2, the testable restrictions in the case of stratified or block randomized trials (Proposition 6) are conditional versions of those in the case of completely randomized trials (Proposition 5). Thus, in what follows we lay out the conditional versions of the null hypotheses, the distributional statistics, and the invariant transformations presented in Section 3A.1.

We first consider the restriction in Proposition 6(a.ii), which yields the following null hypothesis

$$H_0^{1,\mathcal{S}} : F_{Y_{i0}|T_i=0,S_i=s,R_i=r} = F_{Y_{i0}|T_i=1,S_i=s,R_i=r} \text{ for } r = 0,1,\ s \in \mathcal{S}. \tag{3A.5}$$

To obtain the test statistics for the joint hypothesis $H_0^{1,\mathcal{S}}$, we first construct test statistics for a given $s \in \mathcal{S}$,

$$T_{n,m,s}^{1,\mathcal{S}} = \max_{r=0,1} \left\| \sqrt{n}\left(F_{n,Y_{i0}|T_i=0,S_i=s,R_i=r} - F_{n,Y_{i0}|T_i=1,S_i=s,R_i=r}\right)\right\|,$$

$$T^{1,\mathcal{S}}_{n,p,s} = \sum_{r=0,1} p^{r|s}_n \left\| \sqrt{n} \left(F_{n,Y_{i0}|T_i=0,S_i=s,R_i=r} - F_{n,Y_{i0}|T_i=1,S_i=s,R_i=r} \right) \right\|,$$

where $p^{r|s}_n = \sum_{i=1}^n 1\{R_i = r, S_i = s\} / \sum_{i=1}^n 1\{S_i = s\}$. We then aggregate over each of those statistics to get

$$\begin{aligned} T^{1,\mathcal{S}}_{n,m} &= \max_{s \in \mathcal{S}} T^{1,\mathcal{S}}_{n,m,s}, \\ T^{1,\mathcal{S}}_{n,p} &= \sum_{s \in \mathcal{S}} p^s_n T^{1,\mathcal{S}}_{n,p,s}, \text{ where } p^s_n = \sum_{i=1}^n 1\{S_i = s\}/n \text{ for } s \in \mathcal{S}. \end{aligned}$$

In this case, the invariant transformations under $H^{1,\mathcal{S}}_0$ are the ones where n elements are permuted within response-strata subgroups. Formally, for $g \in \mathcal{G}^{1,\mathcal{S}}_0$, $g(\mathbf{Z}) = \{(Y_{i0}, T_{g(i)}, S_{g(i)}, R_{g(i)}) : S_{g(i)} = S_i, R_{g(i)} = R_i, 1 \leq i \leq n\}$, where $\mathbf{Z} = \{(Y_{i0}, T_i, S_i, R_i) : 1 \leq i \leq n\}$. Under $H^{1,\mathcal{S}}_0$ and the cross-sectional i.i.d. assumption within strata, $\mathbf{Z} \overset{d}{=} g(\mathbf{Z})$ for $g \in \mathcal{G}^{1,\mathcal{S}}_0$. Hence, using $\mathcal{G}^{1,S}_0$, we can obtain p-values for $T^{1,\mathcal{S}}_{n,m}$ and $T^{1,\mathcal{S}}_{n,p}$ under $H^{1,\mathcal{S}}_0$.

We now consider testing the restriction in Proposition 6(b.ii). The resulting null hypothesis is given wlog by the following

$$H^{2,\mathcal{S}}_0 : F_{Y_{i0}|T_i=\tau_j,S_i=s,R_i=r_j} = F_{Y_{i0}|T_i=\tau_{j+1},S_i=s,R_i=r_{j+1}} \text{ for } j = 1, \ldots, |\mathcal{T} \times \mathcal{R}| - 1,\ s \in \mathcal{S}. \tag{3A.6}$$

To obtain the test statistics for the joint hypothesis $H^{2,\mathcal{S}}_0$, we first construct test statistics for a given $s \in \mathcal{S}$,

$$\begin{aligned} T^{2,\mathcal{S}}_{n,m,s} &= \max_{j=1,\ldots,|\mathcal{T}\times\mathcal{R}|-1} \left\| \sqrt{n} \left(F_{n,Y_{i0}|T_i=\tau_j,S_i=s,R_i=r_j} - F_{n,Y_{i0}|T_i=\tau_{j+1},S_i=s,R_i=r_{j+1}} \right) \right\|, \\ T^{2,\mathcal{S}}_{n,p,s} &= \sum_{j=1}^{|\mathcal{T}\times\mathcal{R}|-1} w_{j,s} \left\| \sqrt{n} \left(F_{n,Y_{i0}|T_i=\tau_j,S_i=s,R_i=r_j} - F_{n,Y_{i0}|T_i=\tau_{j+1},S_i=s,R_i=r_{j+1}} \right) \right\|, \end{aligned}$$

given fixed or random non-negative weights $w_{j,s}$ for $j = 1, \ldots, |\mathcal{T} \times \mathcal{R}| - 1$ and $s \in \mathcal{S}$. We then aggregate over each of those statistics to get

$$\begin{aligned} T^{2,\mathcal{S}}_{n,m} &= \max_{s \in \mathcal{S}} T^{2,\mathcal{S}}_{n,m,s}, \\ T^{2,\mathcal{S}}_{n,p} &= \sum_{s \in \mathcal{S}} w_s T^{2,\mathcal{S}}_{n,p,s}, \end{aligned}$$

given fixed or random non-negative weights w_s for $s \in \mathcal{S}$.

Under the above hypothesis and the cross-sectional i.i.d. assumption within strata, the distribution of the data is invariant to permutations within strata, i.e. for $g \in \mathcal{G}^{2,\mathcal{S}}_0$, $g(\mathbf{Z}) = \{(Y_{i0}, T_{g(i)}, S_{g(i)}, R_{g(i)}) : S_{g(i)} =$

$S_i, 1 \leq i \leq n\}$. Thus, applying Procedure 1 to $T^{2,\mathcal{S}}_{n,m}$ or $T^{2,\mathcal{S}}_{n,p}$ using $\mathcal{G}^{2,\mathcal{S}}_0$ yields approximately exact p-values for these statistics under $H^{2,\mathcal{S}}_0$.

In practice, it may be possible that response problems could lead to violations of internal validity in some strata but not in others. If that is the case, it may be more appropriate to test interval validity for each stratum separately. Recall that when the goal is to test the IV-R assumption, the stratum-specific hypothesis is $H^{1,s}_0 : F_{Y_{i0}|T_i=0,S_i=s,R_i=r} = F_{Y_{i0}|T_i=1,S_i=s,R_i=r}$ for $r = 0, 1$. Hence, for each $s \in \mathcal{S}$, one can use $\mathcal{G}^{1,\mathcal{S}}_0$ in the above procedure to obtain p-values for $T^{1,\mathcal{S}}_{n,m,s}$ and $T^{1,\mathcal{S}}_{n,p,s}$, and then perform a multiple testing correction that controls either family-wise error rate or false discovery rate. We can follow a similar approach when the goal is to test the IV-P assumption conditional on stratum.

The aforementioned subgroup-randomization procedures split the original sample into respondents and attritors or four treatment-response groups. This approach does not directly extend to cluster randomized experiments.[53] Given the widespread use of regression-based tests in the empirical literature, we illustrate how to test the mean implications of the distributional restrictions of the IV-R and IV-P assumptions using regressions for completely, cluster, and stratified randomized experiments in Section 3B.

3B Regression Tests of Internal Validity

In this section, we show how to implement the mean IV-R and IV-P tests using regression-based procedures. In completely and cluster randomized experiments, the null hypothesis of the IV-R test ($H^1_{0,\mathcal{M}}$) consists of the equality of means across treatment and control responders as well as treatment and control attritors. Meanwhile, the null hypothesis of the IV-P test ($H^2_{0,\mathcal{M}}$) consists of the equality of means across all treatment/respondent subgroups. In the stratified randomization case, the null hypotheses of the IV-R and IV-P tests consist of analogous restrictions *within* strata, $H^{1,\mathcal{S}}_{0,\mathcal{M}}$ and $H^{2,\mathcal{S}}_{0,\mathcal{M}}$, respectively. Here, we present these hypotheses as joint restrictions on linear regression coefficients, which are straightforward to test using the appropriate standard errors. The Stata ado file to implement those regression-based tests is available at https://github.com/daghanem/ATTRITIONTESTS.

[53] To test the distributional restrictions for cluster randomized experiments, the bootstrap-adjusted critical values for the KS and CM-type statistics in Ghanem (2017) can be implemented.

3B.1 Completely and Cluster Randomized Experiments

If the experiment is completely or cluster randomized and Y_{i0} is the baseline outcome, the practitioner may implement one of two equivalent approaches to conducting the mean tests. The first approach is given by:

$$\begin{aligned} Y_{i0} &= \gamma_{11} T_i R_i + \gamma_{01}(1-T_i)R_i + \gamma_{10} T_i (1-R_i) + \gamma_{00}(1-T_i)(1-R_i) + \epsilon_i \\ H^1_{0,\mathcal{M}} &: \gamma_{11} = \gamma_{01} \ \& \ \gamma_{10} = \gamma_{00}, \\ H^2_{0,\mathcal{M}} &: \gamma_{11} = \gamma_{01} = \gamma_{10} = \gamma_{00}. \end{aligned}$$

The second approach allows for an intercept in the regression, which captures the mean baseline outcome for the control attritors:

$$\begin{aligned} Y_{i0} &= \alpha + \beta_{01} R_i + \beta_{10} T_i + \beta_{11} T_i R_i + \epsilon_i \\ H^1_{0,\mathcal{M}} &: \beta_{10} = \beta_{11} = 0, \\ H^2_{0,\mathcal{M}} &: \beta_{01} = \beta_{10} = \beta_{11} = 0. \end{aligned}$$

In some cases, the practitioner may have collected baseline data on determinants of (or proxies for) the outcome of interest, W_{i0} (as defined in Equation 3.3.10). If the practitioner chooses to include these determinants in testing for attrition bias, the regression-based procedure should test the joint hypotheses across the baseline outcome (if available) and the d_W baseline covariates that are relevant for such outcome, i.e. $Z_{i0} = (Y_{i0}, W'_{i0})'$, $\forall j = 1, \ldots, (d_W + 1)$.

$$\begin{aligned} Z^j_{i0} &= \gamma^j_{11} T_i R_i + \gamma^j_{01}(1-T_i)R_i + \gamma^j_{10} T_i (1-R_i) + \gamma^j_{00}(1-T_i)(1-R_i) + \epsilon_i \\ H^1_{0,\mathcal{M}} &: \gamma^j_{11} = \gamma^j_{01} \ \& \ \gamma^j_{10} = \gamma^j_{00} \quad \forall \quad j = 1, \ldots, (d_W + 1) \\ H^2_{0,\mathcal{M}} &: \gamma^j_{11} = \gamma^j_{01} = \gamma^j_{10} = \gamma^j_{00} \quad \forall \quad j = 1, \ldots, (d_W + 1) \end{aligned}$$

As in the univariate case above, the null hypotheses in this multivariate case can also be tested using the specification that inlcudes an intercept. Note that if the researcher is interested instead in testing across multiple *outcomes* we recommend testing these individually rather than jointly (as in Section 3.3.1), while accounting for multiple testing.

3B.2 Stratified Randomized Experiments

As in Section 3B.1, we again present two equivalent formulations of the tests for stratified experiments. In these fully saturated models, the null hypotheses test the equality of means *within* strata. The first version of the test is given by:

$$Y_{i0} = \sum_{s \in \mathcal{S}} \left[\gamma_{11}^s T_i R_i + \gamma_{10}^s T_i (1 - R_i) + \gamma_{01}^s (1 - T_i) R_i + \gamma_{00}^s (1 - T_i)(1 - R_i)\right] 1\{S_i = s\} + \epsilon_i$$

Hence, for $s \in \mathcal{S}$,

$$\begin{aligned} H_{0,\mathcal{M}}^{1,\mathcal{S}} &: \gamma_{11}^s = \gamma_{01}^s \ \& \ \gamma_{10}^s = \gamma_{00}^s, \text{ for all } s \in \mathcal{S}, \\ H_{0,\mathcal{M}}^{2,\mathcal{S}} &: \gamma_{11}^s = \gamma_{01}^s = \gamma_{10}^s = \gamma_{00}^s, \text{ for all } s \in \mathcal{S}. \end{aligned}$$

In this case, the equivalent formulation uses a model with strata fixed effects and strata-specific coefficients,

$$\begin{aligned} Y_{i0} &= \sum_{s=1}^{S} \left\{\alpha^s + \beta_{01}^s R_i + \beta_{10}^s T_i + \beta_{11}^s T_i R_i\right\} 1\{S_i = s\} + \epsilon_i \\ H_{0,\mathcal{M}}^{1,\mathcal{S}} &: \beta_{10}^s = \beta_{11}^s = 0, \text{for all } s \in \mathcal{S}, \\ H_{0,\mathcal{M}}^{2,\mathcal{S}} &: \beta_{01}^s = \beta_{10}^s = \beta_{11}^s = 0, \text{for all } s \in \mathcal{S}. \end{aligned}$$

When the number of strata is large, however, testing the equality of means across groups *within* each stratum may result in high-dimensional inference issues. In that case, practitioners can instead test implications of $H_{0,\mathcal{M}}^{1,\mathcal{S}}$ and $H_{0,\mathcal{M}}^{2,\mathcal{S}}$ as follows:

$$\begin{aligned} Y_{i0} &= \sum_{s=1}^{S} (\alpha^s + \beta_{01}^s R_i) 1\{S_i = s\} + \pi_{10} T_i + \pi_{11} T_i R_i + \epsilon_i \\ H_{0,\mathcal{M}}^{1',\mathcal{S}} &: \pi_{10} = \pi_{11} = 0, \\ Y_{i0} &= \sum_{s=1}^{S} \alpha^s 1\{S_i = s\} + \pi_{01} R_i + \pi_{10} T_i + \pi_{11} T_i R_i + \epsilon_i \\ H_{0,\mathcal{M}}^{2',\mathcal{S}} &: \pi_{01} = \pi_{10} = \pi_{11} = 0. \end{aligned}$$

If the practitioner chooses to include baseline covariates for a stratified experiment, as in Section 3B.1, she should test the joint hypotheses across the baseline outcome and all relevant baseline covariates.

3C Proofs

Proof. (Proposition 5)

(a) Under the assumptions imposed it follows that $F_{U_{i0},U_{i1}|T_i,R_i} = F_{U_{i0},U_{i1}|R_i}$, which implies that for $d = 0, 1$, $F_{Y_{it}(d)|T_i,R_i} = \int 1\{\mu_t(d,u) \leq .\} dF_{U_{it}|T_i,R_i}(u) = \int 1\{\mu_t(d,u) \leq .\} dF_{U_{it}|R_i}(u) = F_{Y_{it}(d)|R_i}$ for $t = 0, 1$. (i) follows by letting $t = 1$ and $d = 0$, while conditioning the left-hand side of the last equation on $T_i = 0$ and $R_i = 1$, and the testable implication in (ii) follows by letting $t = d = 0$.

Following Hsu et al. (2019), we show that the testable restriction is sharp by showing that if $(Y_{i0}, Y_{i1}, T_i, R_i)$ satisfy $Y_{i0}|T_i = 0, R_i = r \overset{d}{=} Y_{i0}|T_i = 1, R_i = r$ for $r = 0, 1$, then there exists (U_{i0}, U_{i1}) such that $Y_{it}(d) = \mu_t(d, U_{it})$ for some $\mu_t(d, .)$ for $d = 0, 1$ and $t = 0, 1$, and $(U_{i0}, U_{i1}) \perp T_i|R_i$ that generate the observed distributions. By the arbitrariness of U_{it} and μ_t, we can let $U_{it} = (Y_{it}(0), Y_{it}(1))'$ and $\mu_t(d, U_{it}) = dY_{it}(1) + (1 - d)Y_{it}(0)$ for $d = 0, 1$, $t = 0, 1$. Note that $Y_{i0} = Y_{i0}(0)$ since $D_{i0} = 0$ w.p.1. Now we need to construct a distribution of $U_i = (U'_{i0}, U'_{i1})$ that satisfies

$$F_{U_i|T_i,R_i} \equiv F_{Y_{i0}(0),Y_{i0}(1),Y_{i1}(0),Y_{i1}(1)|T_i,R_i} = F_{Y_{i0}(0),Y_{i0}(1),Y_{i1}(0),Y_{i1}(1)|R_i}$$

as well as the relevant equalities between potential and observed outcomes. We proceed by first constructing the unobservable distribution for the respondents. By setting the appropriate potential outcomes to their observed counterparts, we obtain the following equalities for the distribution of U_i for the treatment and control respondents

$$F_{U_i|T_i=0,R_i=1} = F_{Y_{i0}(0),Y_{i0}(1),Y_{i1}(0),Y_{i1}(1)|T_i=0,R_i=1} = F_{Y_{i0}(1),Y_{i1},Y_{i1}(1)|Y_{i0},T_i=0,R_i=1} F_{Y_{i0}|T_i=0,R_i=1}$$
$$F_{U_i|T_i=1,R_i=1} = F_{Y_{i0}(1),Y_{i1}(0),Y_{i1}|Y_{i0},T_i=1,R_i=1} F_{Y_{i0}|T_i=1,R_i=1}$$

By construction, $F_{Y_{i0}|T_i,R_i=1} = F_{Y_{i0}|R_i=1}$. Now generating the two distributions above using $F_{Y_{i0}(1),Y_{i1}(0),Y_{i1}(1)|Y_{i0},T_i,R_i=1}$ which satisfies $F_{Y_{i0}(1),Y_{i1},Y_{i1}(1)|Y_{i0},T_i=0,R_i=1} = F_{Y_{i0}(1),Y_{i1}(0),Y_{i1}|Y_{i0},T_i=1,R_i=1}$ yields $U_i \perp T_i|R_i = 1$ and we can construct the observed outcome distribution $(Y_{i0}, Y_{i1})|R_i = 1$ from $U_i|R_i = 1$.

The result for the attritor subpopulation follows trivially from the above arguments,

$$F_{U_i|T_i=0,R_i=0} = F_{Y_{i0}(1),Y_{i1}(0),Y_{i1}(1)|Y_{i0},T_i=0,R_i=0} F_{Y_{i0}|T_i=0,R_i=0},$$
$$F_{U_i|T_i=1,R_i=0} = F_{Y_{i0}(1),Y_{i1}(0),Y_{i1}(1)|Y_{i0},T_i=1,R_i=0} F_{Y_{i0}|T_i=1,R_i=0},$$

Since $F_{Y_{i0}|T_i,R_i=0} = F_{Y_{i0}|R_i=0}$ by construction, it remains to generate the two distributions above using the same $F_{Y_{i0}(1),Y_{i1}(0),Y_{i1}(1)|Y_{i0},R_i=0}$. This leads to a distribution of $U_i|R_i = 0$ that is independent of T_i and that generates the observed outcome distribution $Y_{i0}|R_i = 0$.

(b) Under the given assumptions, it follows that $F_{U_{i0},U_{i1}|T_i,R_i} = F_{U_{i0},U_{i1}|T_i} = F_{U_{i0},U_{i1}}$ where the last equality follows by random assignment. Similar to (a), the above implies that for $d = 0, 1$ and $t = 0, 1$, $F_{Y_{it}(d)|T_i,R_i} = \int 1\{\mu_t(d,u) \leq .\} dF_{U_{it}|T_i,R_i}(u) = \int 1\{\mu_t(d,u) \leq .\} dF_{U_{it}}(u) = F_{Y_{it}(d)}$. (i) follows by letting $t = 1$, while conditioning the left-hand side of the last equation on $T_i = \tau$ and $R_i = 1$ for $d = \tau$ and $\tau = 0, 1$, whereas (ii) follows by letting $d = t = 0$ while conditioning on $T_i = \tau$ and $R_i = r$ for $\tau = 0, 1$, $r = 0, 1$.

To show that the testable restriction is sharp, it remains to show that if $(Y_{i0}, Y_{i1}, T_i, R_i)$ satisfies $Y_{i0}|T_i, R_i \overset{d}{=} Y_{i0}(0)$, then there exists (U_{i0}, U_{i1}) such that $Y_{it}(d) = \mu_t(d, U_{it})$ for some $\mu_t(d, .)$ for $d = 0, 1$ and $t = 0, 1$, and $(U_{i0}, U_{i1}) \perp (T_i, R_i)$. Similar to (a.ii), we let $U_{it} = (Y_{it}(0), Y_{it}(1))'$ and $\mu_t(d, U_{it}) = dY_{it}(1) + (1 - d)Y_{it}(0)$. Then $Y_{i0} = Y_{i0}(0)$ by similar arguments as in the above. Furthermore, $F_{Y_{i0}|T_i,R_i} = F_{Y_{i0}}$ by construction and it follows immediately that

$$
\begin{aligned}
F_{U_i|T_i=0,R_i=1} &= F_{Y_{i0}(1),Y_{i1},Y_{i1}(1)|Y_{i0}T_i=0,R_i=1} F_{Y_{i0}}, \\
F_{U_i|T_i=1,R_i=1} &= F_{Y_{i0}(1),Y_{i1}(0),Y_{i1}|Y_{i0},T_i=1,R_i=1} F_{Y_{i0}}, \\
F_{U_i|T_i=0,R_i=0} &= F_{Y_{i0}(1),Y_{i1}(0),Y_{i1}(1)|Y_{i0},T_i=0,R_i=0} F_{Y_{i0}}, \\
F_{U_i|T_i=1,R_i=0} &= F_{Y_{i0}(1),Y_{i1}(0),Y_{i1}(1)|Y_{i0},T_i=1,R_i=0} F_{Y_{i0}}.
\end{aligned}
$$

Now constructing all of the above distributions using the same $F_{Y_{i0}(1),Y_{i1}(0),Y_{i1}(1)|T_i,R_i}$ that satisfies $F_{Y_{i0}(1),Y_{i1},Y_{i1}(1)|Y_{i0},T_i=0,R_i=1} = F_{Y_{i0}(1),Y_{i1}(0),Y_{i1}|Y_{i0},T_i=1,R_i=1}$ implies the result. □

Proof. (Proposition 6) The proof is immediate from the proof of Proposition 5 by conditioning all statements on S_i. □

Proof. (Proposition 7) For notational brevity, let $U_i = (U_{i0}', U_{i1}')$. We first note that by random assignment, it follows that

$$
F_{U_i|T_i,R_i(0),R_i(1)} = F_{U_i|T_i,\xi(0,V_i),\xi(1,V_i)} = F_{U_i|\xi(0,V_i),\xi(1,V_i)} = F_{U_i|R_i(0),R_i(1)}. \tag{3C.1}
$$

As a result,

$$
F_{U_i|T_i=1,R_i=1} = \frac{p_{01} F_{U_i|(R_i(0),R_i(1))=(0,1)} + p_{11} F_{U_i|(R_i(0),R_i(1))=(1,1)}}{P(R_i = 1|T_i = 1)}, \tag{3C.2}
$$

$$
F_{U_i|T_i=0,R_i=1} = \frac{p_{10} F_{U_i|(R_i(0),R_i(1))=(1,0)} + p_{11} F_{U_i|(R_i(0),R_i(1))=(1,1)}}{P(R_i = 1|T_i = 0)}. \tag{3C.3}
$$

If (i) holds, then $F_{U_i|R_i(0),R_i(1)} = F_{U_i}$, hence

$$F_{U_i|T_i=1,R_i=1} = \frac{p_{01}F_{U_i} + p_{11}F_{U_i}}{P(R_i=1|T_i=1)} = F_{U_i}, \;\; F_{U_i|T_i=0,R_i=1} = \frac{p_{10}F_{U_i} + p_{11}F_{U_i}}{P(R_i=1|T_i=0)} = F_{U_i}.$$

We can similarly show that $F_{U_i|T_i,R_i=0} = F_{U_i}$, it follows trivially that $U_i|T_i, R_i \stackrel{d}{=} U_i|R_i$.

Alternatively, if we assume (ii), $R_i(0) \leq R_i(1)$ implies $p_{10} = 0$. As a result, $P(R_i = 0|T_i = 1) = P(R_i = 0|T_i = 0)$ iff $p_{01} = 0$. It follows that the terms in (3C.2) and (3C.3) both equal $F_{U_i|(R_i(0),R_i(1))=(1,1)}$. Similarly, it follows that $F_{U_i|T_i=1,R_i=0} = F_{U_i|T_i=0,R_i=0} = F_{U_i|(R_i(0),R_i(1))=(0,0)}$, which implies the result.

Finally, suppose (iii) holds, then equal attrition rates imply that $p_{01} = p_{10}$. The exchangeability restriction implies that $F_{U_i|(R_i(0),R_i(1))=(0,1)} = F_{U_i|(R_i(0),R_i(1))=(1,0)}$. Hence,

$$\begin{aligned} F_{U_i|T_i=1,R_i=1} &= \frac{p_{01}F_{U_i|(R_i(0),R_i(1))=(0,1)} + p_{11}F_{U_i|(R_i(0),R_i(1))=(1,1)}}{P(R_i=1|T_i=1)} \\ &= \frac{p_{10}F_{U_i|(R_i(0),R_i(1))=(1,0)} + p_{11}F_{U_i|(R_i(0),R_i(1))=(1,1)}}{P(R_i=1|T_i=0)} = F_{U_i|T_i=0,R_i=1}. \end{aligned} \tag{3C.4}$$

Similarly, it follows that $F_{U_i|T_i=1,R_i=0} = F_{U_i|T_i=0,R_i=0}$, which implies the result. □

3C.1 Supplementary Example for Section 3.3.4

Suppose that there are two unobservables that enter the outcome equation, $U_{it} = (U_{it}^1, U_{it}^2)'$ for $t = 0, 1$, such that $(U_{i0}^1, U_{i1}^1) \perp T_i|R_i$ whereas $(U_{i0}^2, U_{i1}^2) \not\perp T_i|R_i$. Let the outcome at baseline be a trivial function of U_{i0}^2, whereas the outcome in the follow-up period is a non-trivial function of both U_{i0}^1 and U_{i0}^2, e.g.

$$\begin{aligned} Y_{i0} &= U_{i0}^1 \\ Y_{i1} &= U_{i1}^1 + U_{i1}^2 + T_i(\beta_1 U_{i1}^1 + \beta_2 U_{i1}^2) \end{aligned}$$

As a result, even though $Y_{i0}|T_i = 1, R_i \stackrel{d}{=} Y_{i0}|T_i = 0, R_i$ holds, $Y_{i1}(0)|T_i = 1, R_i = 1 \stackrel{d}{\neq} Y_{i1}|T_i = 0, R_i = 1$. In other words, the control respondents do not provide a valid counterfactual for the treatment respondents in the follow-up period despite the identity of the baseline outcome distribution for treatment and control groups conditional on response status. We can illustrate this by looking at the average treatment effect for the treatment respondents,

$$\begin{aligned} &E[Y_{i1}(1) - Y_{i1}(0)|T_i = 1, R_i = 1] \\ &= \underbrace{E[U_{i1}^1 + U_{i1}^2 + \beta_1 U_{i1}^1 + \beta_2 U_{i1}^2|T_i = 1, R_i = 1]}_{E[Y_{i1}|T_i=1,R_i=1]} - \underbrace{E[U_{i1}^1 + U_{i1}^2|T_i = 1, R_i = 1]}_{\neq E[Y_{i1}|T_i=0,R_i=1]}. \end{aligned}$$

Hence, $E[Y_{i1}|T_i = 1, R_i = 1] - E[Y_{i1}|T_i = 0, R_i = 1] \neq \beta_1 E[U^1_{i1}|T_i = 1, R_i = 1] + \beta_2 E[U^2_{i1}|T_i = 1, R_i = 1]$, i.e. the difference in mean outcomes between treatment and control respondents does not identify an average treatment effect for the treatment respondents.

We could however have a case in which the control respondents provide a valid counterfactual for the treatment respondents even though the treatment effect for individual i depends on an unobservable that is not independent of treatment conditional on response, i.e. U^2_{it}. Specifically, let

$$Y_{it} = U^1_{it} + T_i(\beta_1 U^1_{it} + \beta_2 U^2_{it}) \tag{3C.5}$$

and consider the identification of an average treatment effect, $E[Y_{i1}(1) - Y_{i1}(0)|T_i = 1, R_i = 1] = E[U^1_{i1} + \beta_1 U^1_{i1} + \beta_2 U^2_{i1}|T_i = 1, R_i = 1] - E[U^1_{i1}|T_i = 1, R_i = 1] = E[Y_{i1}|T_i = 1, R_i = 1] - E[Y_{i1}|T_i = 0, R_i = 1]$, since $E[U^1_{i1}|T_i = 1, R_i = 1] = E[U^1_{i1}|T_i = 0, R_i = 1]$. Note however that in this case what we identify is no longer internally valid for the entire respondent subpopulation, but for the smaller subpopulation of treatment respondents.

3D Selection of Articles from the Field Experiment Literature

3D.1 Selection of Articles for the Review

In order to understand both the extent of attrition as well as how authors test for attrition bias in practice, we systematically reviewed articles that report the results of field experiments. We include articles that were published in the top five journals in economics, as well as five highly regarded applied economics journals: *American Economic Review*, *American Economic Journal: Applied Economics*, *Econometrica*, *Economic Journal*, *Journal of Development Economics*, *Journal of Human Resources*, *Journal of Political Economy*, *Review of Economics and Statistics*, *Review of Economic Studies*, and *Quarterly Journal of Economics*.[54] By searching for *RCT*, *randomized controlled trial*, or *field experiment* in each journal's website, we identified 202 articles that were published between 2009 and 2015.[55] From these 202 articles, we review those in which the main goal is to report the results of a field experiment and for which attrition is relevant given the experiment's study design. To be consistent with our panel approach in Section 3.3 in the paper, we only focus on those experiments with baseline data on at least one main outcome variable.

Table 3A displays the distribution of the 93 articles that satisfied the selection criteria by journal and year

[54] We chose these four applied journals because they are important sources of published field experiments.

[55] Our initial search using these keywords yielded 235 articles but 33 of those papers were excluded since they were observational studies exploiting some sort of quasi-experimental variation.

of publication.[56] Of these 93 articles, 61% were published in the *Journal of Development Economics*, the *American Economic Journal: Applied Economics*, and the *Quarterly Journal of Economics*. Approximately 56% of our sample of articles were published in 2014 and 2015.

Table 3A: Distribution of Articles by Journal and Year of Publication

Journal	Year							Total
	2009	2010	2011	2012	2013	2014	2015	
AEJ: Applied	0	0	0	3	3	3	8	17
AER	0	1	1	2	0	2	2	8
EJ	0	0	1	2	0	5	0	8
Econometrica	1	0	0	0	0	1	0	2
JDE	0	0	1	1	3	11	6	22
JHR	0	0	0	1	1	1	2	5
JPE	0	0	1	0	0	0	0	1
QJE	1	1	4	3	2	4	3	18
REstat	2	0	2	1	1	1	3	10
REstud	0	0	0	0	1	1	0	2
Total	4	2	10	13	11	29	24	93

Notes: The 93 articles that we include in our review correspond to 96 field experiments. The two articles that reported more than one field experiment are published in the AER(2015) and the QJE(2011), respectively.

We also exclude 64 articles that do not have available baseline data for any of the outcomes reported in the abstract. From these papers, 52% do not collect baseline outcome and 5% collect baseline data but have a baseline attrition above fifty percent. The remaining 28 papers that we exclude (43%) have the same baseline outcome for everyone by design. Some examples in this category include training interventions that target unemployed individuals and measure impacts on employment, and interventions that aim to estimate which of the multiple treatment arms has a higher impact on the take-up of a newly introduced product.

One challenge that arose in our review was determining which attrition rates and attrition tests are most relevant, since the reported attrition rates usually vary across different data sources or different subsamples. We chose to focus on the results that are reported in the abstract in our analysis of attrition rates. But, since many authors do not report attrition tests for each of the abstract results, in our analysis of attrition tests we focus on whether authors report a test that is relevant to at least one abstract result.

3D.2 Selection of Articles for the Empirical Applications

In order to conduct the empirical applications in Section 3.5, we identified 47 articles that had publicly available analysis files from the 93 articles in our review (see Section 3.2). To select the five articles that had the highest attrition rates from that group, we reviewed the data files for twelve articles. We excluded

[56] Some of the articles report results for more than one intervention. Thus, these 93 articles correspond to 96 field experiments.

field experiments for a variety of reasons that would not, in the majority of cases, affect the ability of the authors to implement our tests. Of the seven experiments that were excluded: two did not provide the data sets along with the analysis files due to confidentiality restrictions, two provided the data sets but did not include attritors, and one did not provide sufficient information to identify the attritors. In two cases, an exceptionally high number of missing values at baseline was the limiting factor since the attrition rate at follow-up conditional on baseline response was lower than the attrition rate reported in the paper.

3E Attrition Tests in the Field Experiment Literature

In order to classify the attrition tests that are conducted in the 93 articles that we review, we gathered information on the different econometric strategies that were carried out to test for attrition bias. In this section, we describe these empirical strategies and classify them into differential attrition rates test, selective attrition tests, and determinants of attrition test. We specify the null hypotheses of the selective attrition tests since this test is closely related to the tests that we propose. In contrast, we categorize the estimation strategies for the differential attrition rates test and the determinants of attrition test as broadly as possible and include any article that performs a regression under any of these two categories as performing the relevant test. Throughout this section, we use the following notation to facilitate the exposition of each strategy and the comparison across them:

-Let R_i take the value of 1 if individual i belongs to the follow-up sample.

-Let T_i take the value of 1 if individual i belongs to the treatment group.

-Let X_{i0} be a $k \times 1$ vector of baseline variables.

-Let Y_{i0} be a $l \times 1$ vector of outcomes collected at baseline.

-Let $Z_{i0} = (X'_{i0}, Y'_{i0})'$.

-For a vector w, w^j denotes the j^{th} element of w.

3E.1 Differential Attrition Rates Test

The *differential attrition rates test* determines whether the rates of attrition are statistically significantly different across treatment and control groups.

1. t-test of the equality of attrition rate by treatment group, i.e. $H_0 : P(R_i = 0|T_i = 1) = P(R_i = 0|T_i = 0)$.

2. $R_i = \gamma + T_i\beta + U_i$; may include strata fixed effects.

3. $R_i = \gamma + T_i\beta + X'_{i0}\theta + Y'_{i0}\alpha + U_i$; may include strata fixed effects.

3E.2 Selective Attrition Test

The *selective attrition test* determines whether, conditional on response status, the distribution of observable characteristics is the same across treatment and control groups. We identify two sub-types of selective attrition tests: i) a test that includes only respondents or attritors, and ii) a test that includes both respondents and attritors. We note that the selective attrition tests are usually conducted on both baseline outcomes and baseline covariates. Some authors conduct multiple tests for *individual* baseline variables while others test *all* baseline variables jointly (see Table 3C for details). Thus, for each estimation strategy, we report the null hypotheses that are used in each case.

Tests that include only respondents or attritors

1. t-test of baseline characteristics by treatment group among respondents:

   (a) *Multiple hypotheses for individual baseline variables:*

   For each $j = 1, 2, \ldots, (l + k)$

   $$H_0^j : E[Z_{i0}^j | T_i = 1, R_i = 1] = E[Z_{i0}^j | T_i = 0, R_i = 1].$$

   (b) *Joint hypothesis for all baseline variables:*

   $$H_0 : E[Z_{i0}^j | T_i = 1, R_i = 1] = E[Z_{i0}^j | T_i = 0, R_i = 1], \ \forall j = 1, \ldots, (l + k).$$

2. $T_i = \gamma + X_{i0}'\theta + Y_{i0}'\alpha + U_i$ if $R_i = 1$; may include strata fixed effects.

   (a) *Joint hypothesis for all baseline variables:*

   $$H_0 : \theta = \alpha = 0$$

3. Kolmogorov-Smirnov (KS) test of baseline characteristics by treatment group among respondents.

   (a) *Multiple hypotheses for individual baseline variables:*

   For each $j = 1, 2, \ldots, (l + k)$

   $$H_0^j : F_{Z_{i0}^j | T_i, R_i = 1} = F_{Z_{i0}^j | R_i = 1}$$

4. $Z_{i0}^j = \gamma + T_i \beta^j + U_i^j$ if $R_i = 1$, for $j = 1, 2, \ldots, (l + k)$; may include strata fixed effects.

(a) *Multiple hypotheses for individual baseline variables:*

For each $j = 1, 2, \ldots, (l + k)$

$$H_0^j : \beta^j = 0$$

(b) *Joint hypothesis for all baseline variables:*

$$H_0 : \beta^1 = \beta^2 = \cdots = \beta^{l+k} = 0$$

5. $Z_{i0}^j = \gamma + T_i\beta^j + U_i^j$ if $R_i = 0$, for $j = 1, 2, \ldots, (l + k)$; may include strata fixed effects.

   (a) *Multiple hypotheses for individual baseline variables:*

   For each $j = 1, 2, \ldots, (l + k)$

$$H_0^j : \beta^j = 0$$

Tests that include both respondents and attritors

1. $Z_{i0}^j = \gamma^j + T_i\beta^j + (1 - R_i)\lambda^j + T_i(1 - R_i)\phi^j + U_i^j$ for $j = 1, 2, \ldots, (l + k)$; may include strata fixed effects.

   (a) *Multiple hypotheses for individual baseline variables:*[57]

   For each $j = 1, 2, \ldots, (l + k)$

$$H_0^j : \beta^j = 0$$

2. $R_i = \gamma + T_i\beta + X_{i0}'\theta + Y_{i0}'\alpha + T_iX_{i0}'\lambda_1 + T_iY_{i0}'\lambda_2 + U_i$; may include strata fixed effects.

   (a) *Multiple hypotheses for individual baseline variables I:*

   For each $m = 1, 2, \ldots, k$ and $j = 1, 2, \ldots, l$

$$H_0^{\theta,m} : \theta^m = 0 \quad , \quad H_0^{\alpha,j} : \alpha^j = 0 \quad , \quad H_0^{\lambda_{1,m}} : \lambda_1^m = 0 \quad , \quad H_0^{\lambda_{2,j}} : \lambda_2^j = 0$$

   (b) *Multiple hypotheses for individual baseline variables II:*

[57] Although this null hypothesis is testing for the equality of means for treatment and control respondents, we classify this strategy as one that includes both respondents and attritors given that the regression test is based on both samples.

For each $m = 1, 2, \ldots, k$ and $j = 1, 2, \ldots, l$

$$H_0^{\lambda_1, m} : \lambda_1^m = 0 \quad , \quad H_0^{\lambda_{2,j}} : \lambda_2^j = 0$$

(c) *Joint hypothesis for all baseline variables I:*

$$H_0 : \beta = \theta = \alpha = \lambda_1 = \lambda_2 = 0$$

(d) *Joint hypothesis for all baseline variables II:*

$$H_0 : \lambda_1 = \lambda_2 = 0$$

3. t-test of the equality of the difference in baseline outcome between respondents and attritors across treatment groups.

   (a) *Multiple hypotheses for individual baseline outcomes:*

   For each $j = 1, 2, \ldots, l$

$$\begin{aligned} H_0^j : E[Y_{i0}^j | T_i = 1, R_i = 1] - E[Y_{i0}^j | T_i = 1, R_i = 0] \\ = E[Y_{i0}^j | T_i = 0, R_i = 1] - E[Y_{i0}^j | T_i = 0, R_i = 0] \end{aligned}$$

3E.3 Determinants of Attrition Test

The *determinants of attrition test* determines whether attritors are significantly different from respondents regardless of treatment assignment.

1. $R_i = \gamma + T_i\beta + X_{i0}'\theta + Y_{i0}'\alpha + U_i$; may include strata fixed effects.
2. $Z_{i0}^j = \gamma^j + (1 - R_i)\lambda^j + U_i^j$, $j = 1, 2, \ldots, (l + k)$; may include strata fixed effects.
3. $R_i = \gamma + X_{i0}'\theta + Y_{i0}'\alpha + U_i$; may include strata fixed effects.
4. Let $Reason_i$ take the value of 1 if the individual identifies it as one of the reasons for which she dropped out of the program. The test consists of a Probit estimation of:
   $Reason_i = \gamma + T_i\beta + U_i$ if $R_i = 1$; may include strata fixed effects.

Table 3B: Overall Attrition Rate by Country's Income Group

Field Experiments in:	*N*	*Mean*	*SD*	*Min*	*Max*	*p25*	*p75*	Prop. of Experiments with Rate > 15%
High income countries	28	20.7	24.2	0	87	3	28	46%
Upper middle income countries	18	15.6	13.1	0	54	7	20	55%
Low and lower middle income countries	47	11.9	12.6	0	59	2	18	34%
All countries	93	15.3	17.2	0	87	3.3	21	42%

Notes: This table considers the highest overall attrition rate for each field experiment in our review and excludes one paper that does not report overall attrition rates. We classify countries by income group according to the official definition of the World Bank.

Table 3C: Number of Baseline Variables Included in The Selective Attrition Test

Category	**No. of Baseline Variables Included**					
	Mean	*SD*	*Min*	*Max*	*p25*	*p75*
All papers that conduct a selective attrition test	17.2	10.3	1	46	10	21
Papers that test on multiple baseline variables:						
Multiple hypotheses for individual variables (75%)	16.6	9.7	2	46	10	21
Joint hypothesis for all variables (25%)	20.3	11.3	4	44	13	23

Notes: Of the 46 experiments that conduct a selective attrition test, 44 test on multiple baseline variables. This table excludes one experiment that tests on multiple baseline variables but does not provide sufficient information for it to be categorized. Percentages are a proportion of the 44 experiments that test on multiple baseline variables.

3F Equal Attrition Rates with Multiple Treatment Groups

In this section, we illustrate that once we have more than two treatment groups and violations of monotonicity, then equal attrition rates are possible without imposing the equality of proportions of certain subpopulations unlike Example 2 in the paper. Consider the case where we have three treatment groups, i.e. $T_i \in \{0,1,2\}$. For brevity, we use the notation $P_i((r_0,r_1,r_2)) \equiv P((R_i(0),R_i(1),R_i(2)) = (r_0,r_1,r_2))$ for $(r_0,r_1,r_2) \in \{0,1\}^3$. Hence,

$$\begin{aligned}
P(R_i = 0|T_i = 0) &= P_i((0,0,0)) + P_i((0,0,1)) + P_i((0,1,0)) + P_i((0,1,1)) \\
P(R_i = 0|T_i = 1) &= P_i((0,0,0)) + P_i((0,0,1)) + P_i((1,0,0)) + P_i((1,0,1)) \\
P(R_i = 0|T_i = 2) &= P_i((0,0,0)) + P_i((1,0,0)) + P_i((0,1,0)) + P_i((1,1,0))
\end{aligned} \tag{3F.1}$$

The equality of attrition rates across the three groups, i.e. $P(R_i = 0|T_i = 0) - P(R_i = 0|T_i = 1) = P(R_i = 0|T_i = 0) - P(R_i = 0|T_i = 2) = 0$ implies the following equalities,

$$\begin{aligned} P_i((0,1,0)) + P_i((0,1,1)) &= P_i((1,0,0)) + P_i((1,0,1)) \\ P_i((0,0,1)) + P_i((0,1,1)) &= P_i((1,0,0)) + P_i((1,1,0)) \end{aligned} \tag{3F.2}$$

which can occur without constraining the proportions of different subpopulations to be equal.

3G Identification and Testing for the Multiple Treatment Case

In this section, we present the generalization of Propositions 5 and 6 (Section 3G.1) as well as the distributional test statistics (Section 3G.2) in the paper to the case where the treatment variable has arbitrary finite-support. As in the paper, we provide results for completely and stratified randomized experiments. We maintain that $D_{i0} = 0$ for all i, i.e. no treatment is assigned in the baseline period, $D_{i1} \in \mathcal{D}$, where wlog $\mathcal{D} = \{0, 1, \ldots, |\mathcal{D}| - 1\}$, $|\mathcal{D}| < \infty$. $D_i \equiv (D_{i0}, D_{i1}) \in \{(0,0), (0,1), \ldots, (0, |\mathcal{D}| - 1)\}$. Let T_i denote the indicator for membership in the treatment group defined by D_i, i.e. $T_i \in \mathcal{T} = \{0, 1, \ldots, |\mathcal{D}| - 1\}$, where $T_i = D_{i1}$ and hence $|\mathcal{T}| = |\mathcal{D}|$ by construction.

3G.1 Identification and Sharp Testable Restrictions

Completely Randomized Trials

Proposition 8. *Assume* $(U_{i0}, U_{i1}, V_i) \perp T_i$.

(a) If $(U_{i0}, U_{i1}) \perp T_i|R_i$ *holds, then*

(i) (Identification) $Y_{i1}|T_i = \tau, R_i = 1 \overset{d}{=} Y_{i1}(\tau)|R_i = 1$ *for* $\tau \in \mathcal{T}$.

(ii) (Sharp Testable Restriction) $Y_{i0}|T_i = \tau, R_i = r \overset{d}{=} Y_{i0}|T_i = \tau', R_i = r$ *for* $r = 0, 1$, *for* $\tau, \tau' \in \mathcal{T}, \tau \neq \tau'$.

(b) If $(U_{i0}, U_{i1}) \perp R_i|T_i$ *holds, then*

(i) (Identification) $Y_{i1}|T_i = \tau, R_i = 1 \overset{d}{=} Y_{i1}(\tau)$ *for* $\tau \in \mathcal{T}$.

(ii) (Sharp Testable Restriction) $Y_{i0}|T_i = \tau, R_i = r \overset{d}{=} Y_{i0}$ *for* $\tau \in \mathcal{T}$, $r = 0, 1$.

Proof. (Proposition 8) (a) Under the assumptions imposed it follows that $F_{U_{i0},U_{i1}|T_i,R_i} = F_{U_{i0},U_{i1}|R_i}$, which implies that for $d \in \mathcal{D}$, $F_{Y_{it}(d)|T_i,R_i} = \int 1\{\mu_t(d,u) \leq .\} dF_{U_{it}|T_i,R_i}(u) = \int 1\{\mu_t(d,u) \leq .\} dF_{U_{it}|R_i}(u) =$

$F_{Y_{it}(d)|R_i}$. (i) follows by letting $t = 1$ and $d = \tau$, while conditioning the left-hand side of the last equation on $T_i = \tau$ and $R_i = 1$ and the right-hand side on $R_i = 1$. The testable implication in (ii) follows by letting $t = d = 0$ and conditioning the left-hand side on $T_i = \tau$ and $R_i = r$ and the right-hand side on $T_i = \tau'$ and $R_i = r$, where $\tau \neq \tau'$.

Following Hsu et al. (2019), we show that the testable restriction is sharp by showing that if $(Y_{i0}, Y_{i1}, T_i, R_i)$ satisfy $Y_{i0}|T_i = \tau, R_i = r \overset{d}{=} Y_{i0}|T_i = \tau', R_i = r$ for $r = 0, 1$, $\tau, \tau' \in \mathcal{T}$, $\tau \neq \tau'$, then there exists (U_{i0}, U_{i1}) such that $Y_{it}(d) = \mu_t(d, U_{it})$ for some $\mu_t(d, .)$ for $d \in \mathcal{D}$ and $t = 0, 1$ and $(U_{i0}, U_{i1}) \perp T_i|R_i$ that generate the observed distributions. By the arbitrariness of U_{it} and μ_t, we can let $U'_{it} = \mathbf{Y}_{it}(.) = (Y_{it}(0), Y_{it}(1), \ldots, Y_{it}(|\mathcal{D}| - 1))$ and $\mu_t(d, U_{it}) = \sum_{j=0}^{\mathcal{D}-1} 1\{j = d\}Y_{it}(j)$ for $d \in \mathcal{D}$, $t = 0, 1$. Note that $Y_{i0} = Y_{i0}(0)$ since $D_{i0} = 0$ w.p.1. Now we have to construct a distribution of $U_i = (U'_{i0}, U'_{i1})$ that satisfies

$$F_{U_i|T_i,R_i} \equiv F_{\mathbf{Y}_{i0}(.),\mathbf{Y}_{i1}(.)|T_i,R_i} = F_{\mathbf{Y}_{i0}(.),\mathbf{Y}_{i1}(.)|R_i}$$

as well as the relevant equalities between potential and observed outcomes. We proceed by first constructing the unobservable distribution for the respondents. By setting the appropriate potential outcomes to their observed counterparts, we obtain the following equalities for the distribution of U_i for the respondents in the different treatment groups

$$\begin{aligned} F_{U_i|T_i=\tau,R_i=1} &= F_{\{Y_{i0}(d)\}_{d=1}^{|\mathcal{D}|-1},\mathbf{Y}_{i1}(.)|Y_{i0},T_i=\tau,R_i=1} F_{Y_{i0}|T_i=\tau,R_i=1} \\ &= F_{\{Y_{i0}(d)\}_{d=1}^{|\mathcal{D}|-1},\{Y_{i1}(d)\}_{d=0}^{\tau-1},Y_{i1},\{Y_{i1}(d)\}_{d=\tau+1}^{|\mathcal{D}|-1}|Y_{i0},T_i=\tau,R_i=1} F_{Y_{i0}|T_i=\tau,R_i=1}. \end{aligned} \tag{3G.1}$$

By construction, $F_{Y_{i0}|T_i,R_i=1} = F_{Y_{i0}|R_i=1}$. Now generating the above distribution for all $\tau \in \mathcal{T}$ such that $F_{\{Y_{i0}(d)\}_{d=1}^{|\mathcal{D}|-1},\{Y_{i1}(d)\}_{d=0}^{\tau-1},Y_{i1},\{Y_{i1}(d)\}_{d=\tau+1}^{|\mathcal{D}|-1}|Y_{i0},T_i=\tau,R_i=1}$ which satisfies the following equality $\forall \tau, \tau' \in \mathcal{T}$, $\tau \neq \tau'$,

$$\begin{aligned} &F_{\{Y_{i0}(d)\}_{d=1}^{|\mathcal{D}|-1},\{Y_{i1}(d)\}_{d=0}^{\tau-1},Y_{i1},\{Y_{i1}(d)\}_{d=\tau+1}^{|\mathcal{D}|-1}|Y_{i0},T_i=\tau,R_i=1} \\ =&F_{\{Y_{i0}(d)\}_{d=1}^{|\mathcal{D}|-1},\{Y_{i1}(d)\}_{d=0}^{\tau'-1},Y_{i1},\{Y_{i1}(d)\}_{d=\tau'+1}^{|\mathcal{D}|-1}|Y_{i0},T_i=\tau',R_i=1}, \end{aligned}$$

yields $U_i \perp T_i|R_i = 1$ and we can construct the observed outcome distribution $(Y_{i0}, Y_{i1})|R_i = 1$ from $U_i|R_i = 1$.

The result for the attritor subpopulation follows trivially from the above arguments,

$$F_{U_i|T_i=\tau,R_i=0} = F_{\{Y_{i0}(d)\}_{d=1}^{|\mathcal{D}|-1},\mathbf{Y}_{it}(.)|Y_{i0},T_i=\tau,R_i=0} F_{Y_{i0}|T_i=\tau,R_i=0} \tag{3G.2}$$

Since $F_{Y_{i0}|T_i,R_i=0} = F_{Y_{i0}|R_i=0}$ by construction, it remains to generate the above distribution for all $\tau \in \mathcal{T}$ using the same $F_{\{Y_{i0}(d)\}_{d=1}^{|\mathcal{D}|-1},\mathbf{Y}_{it}(.)|Y_{i0},R_i=0}$. This leads to a distribution of $U_i|R_i = 0$ that is independent of T_i and that generates the observed outcome distribution $Y_{i0}|R_i = 0$.

(b) Under the given assumptions, it follows that $F_{U_{i0},U_{i1}|T_i,R_i} = F_{U_{i0},U_{i1}|T_i} = F_{U_{i0},U_{i1}}$ where the last equality follows by random assignment. Similar to (a), the above implies that for $d \in \mathcal{D}$, $F_{Y_{it}(d)|T_i,R_i}(.) = \int 1\{\mu_t(d,u) \leq .\}dF_{U_{it}|T_i,R_i}(u) = \int 1\{\mu_t(d,u) \leq .\}dF_{U_{it}}(u) = F_{Y_{it}(d)}$. (i) follows by letting $d = \tau$ and $t = 1$, while conditioning the left-hand side of the last equation on $T_i = \tau$ and $R_i = 1$, whereas (ii) follows by letting $d = t = 0$ while conditioning on $T_i = \tau$ and $R_i = r$ for $\tau \in \mathcal{T}$, $r = 0, 1$.

To show that the testable restriction is sharp, it remains to show that if $(Y_{i0}, Y_{i1}, T_i, R_i)$ satisfies $Y_{i0}|T_i, R_i \stackrel{d}{=} Y_{i0}(0)$, then there exists (U_{i0}, U_{i1}) such that $Y_{it}(d) = \mu_t(d, U_{it})$ for some $\mu_t(d,.)$ for $d \in \mathcal{D}$ and $t = 0, 1$ and $(U_{i0}, U_{i1}) \perp (T_i, R_i)$. Similar to (a.ii), we let $U'_{it} = \mathbf{Y}_{it}(.) = (Y_{it}(0), Y_{it}(1), \ldots, Y_{it}(|\mathcal{D}| - 1))$ and $\mu_t(d, U_{it}) = \sum_{j=0}^{\mathcal{D}-1} 1\{j = d\}Y_{it}(j)$ for $d \in \mathcal{D}$, $t = 0, 1$. By construction, $Y_{i0} = Y_{i0}(0)$. Furthermore, $F_{Y_{i0}|T_i,R_i} = F_{Y_{i0}}$ by assumption. It follows immediately that for all $\tau \in \mathcal{T}$

$$F_{U_i|T_i=\tau,R_i=1} = F_{\{Y_{i0}(d)\}_{d=1}^{|\mathcal{D}|-1},\{Y_{i1}(d)\}_{d=0}^{\tau-1},Y_{i1},\{Y_{i1}(d)\}_{d=\tau+1}^{|\mathcal{D}|-1}|T_i=\tau,R_i=1}F_{Y_{i0}},$$
$$F_{U_i|T_i=\tau,R_i=0} = F_{\{Y_{i0}(d)\}_{d=1}^{|\mathcal{D}|-1},\mathbf{Y}_{it}(.)|Y_{i0},T_i=\tau,R_i=0}F_{Y_{i0}}.$$

Now constructing all of the above distributions using the same $F_{\{Y_{i0}(d)\}_{d=1}^{|\mathcal{D}|-1},\mathbf{Y}_{it}(.)|Y_{i0},T_i,R_i}$ that satisfies the above equalities for all $\tau \in \mathcal{T}$ implies the result. □

Stratified Randomized Trials

Proposition 9. *Assume* $(U_{i0}, U_{i1}, V_i) \perp T_i|S_i$.

(a) If $(U_{i0}, U_{i1}) \perp T_i|S_i, R_i$ *holds, then*

- *(i) (Identification)* $Y_{i1}|T_i = \tau, S_i = s, R_i = 1 \stackrel{d}{=} Y_{i1}(\tau)|S_i = s, R_i = 1$, *for* $\tau \in \mathcal{T}, s \in \mathcal{S}$.
- *(ii) (Sharp Testable Restriction)* $Y_{i0}|T_i = \tau, S_i = s, R_i = r \stackrel{d}{=} Y_{i0}|T_i = \tau', S_i = s, R_i = r$, $\forall \tau, \tau' \in \mathcal{T}, \tau \neq \tau, s \in \mathcal{S}, r = 0, 1$.

(b) If $(U_{i0}, U_{i1}) \perp R_i|T_i$ *holds, then*

- *(i) (Identification)* $Y_{i1}|T_i = \tau, S_i = s, R_i = 1 \stackrel{d}{=} Y_{i1}(\tau)|S_i = s$ *for* $\tau \in \mathcal{T}$, $s \in \mathcal{S}$.
- *(ii) (Sharp Testable Restriction)* $Y_{i0}|T_i = \tau, S_i = s, R_i = r \stackrel{d}{=} Y_{i0}|S_i = s$ *for* $\tau \in \mathcal{T}$, $r = 0, 1$, $s \in \mathcal{S}$.

Proof. (Proposition 9) The proof for this proposition follows in a straightforward manner from the proof for Proposition 8 by conditioning all statements on S_i. □

3G.2 Distributional Test Statistics

Next, we present the null hypotheses and distributional statistics for the multiple treatment case. For simplicity, we only present the joint statistics that take the maximum to aggregate over the individual statistics of each distributional equality implied by a given testable restriction.

Completely Randomized Trials

The null hypothesis implied by Proposition 8(a.ii) is given by the following,

$$H_0^{1,\mathcal{T}} : F_{Y_{i0}|T_i=\tau,R_i=r} = F_{Y_{i0}|T_i=\tau',R_i=r} \text{ for } \tau,\tau' \in \mathcal{T},\ \tau \neq \tau',\ r = 0,1. \tag{3G.3}$$

Consider the following general form of the distributional statistic for the above null hypothesis is $T_n^{1,\mathcal{T}} = \max_{r\in\{0,1\}} T_{n,r}^{1,\mathcal{T}}$, where for $r = 0, 1$,

$$T_{n,r}^{1,\mathcal{T}} = \max_{(\tau,\tau')\in\mathcal{T}^2:\tau\neq\tau'} \left\| \sqrt{n} \left(F_{n,Y_{i0}|T_i=\tau,R_i=r} - F_{n,Y_{i0}|T_i=\tau',R_i=r} \right) \right\|.$$

The randomization procedure proposed in the paper using the transformations $\mathcal{G}_0^1$ can be used to obtain p-values for the above statistic under $H_0^{1,\mathcal{T}}$.

Let $(\tau, r) \in \mathcal{T} \times \mathcal{R}$, where $\mathcal{R} = \{0, 1\}$. Let (τ_j, r_j) denote the j^{th} element of $\mathcal{T} \times \mathcal{R}$, then the null hypothesis implied by Proposition 8(b.ii) is given by the following:

$$H_0^{2,\mathcal{T}} : F_{Y_{i0}|T_i=\tau_j,R_i=r_j} = F_{Y_{i0}|T_i=\tau_{j+1},R_i=r_{j+1}} \text{ for } j = 1,\ldots,|\mathcal{T}\times\mathcal{R}| - 1. \tag{3G.4}$$

the test statistic for the above *joint* hypothesis is given by

$$T_{n,m}^{2,\mathcal{T}} = \max_{j=1,\ldots,|\mathcal{T}\times\mathcal{R}|-1} \left\| \sqrt{n} \left(F_{n,Y_{i0}|T_i=\tau_j,R_i=r_j} - F_{n,Y_{i0}|T_i=\tau_{j+1},R_i=r_{j+1}} \right) \right\|,$$

The randomization procedure proposed in the paper using the transformations $\mathcal{G}_0^2$ can be used to obtain p-values for the above statistic under $H_0^{2,\mathcal{T}}$.

Stratified Randomized Trials

The null hypothesis implied by Proposition 9(a.ii) is given by the following,

$$H_0^{1,\mathcal{S},\mathcal{T}} : F_{Y_{i0}|T_i=\tau,S_i=s,R_i=r} = F_{Y_{i0}|T_i=\tau',S_i=s,R_i=r} \text{ for } \tau,\tau' \in \mathcal{T},\ \tau \neq \tau',\ s \in \mathcal{S},\ r = 0,1. \tag{3G.5}$$

Consider the following general form of the distributional statistic for the above null hypothesis is $T_n^{1,\mathcal{S},\mathcal{T}} = \max_{s\in\mathcal{S}} \max_{r\in\{0,1\}} T_{n,r,s}^{1,\mathcal{T}}$, where for $s \in \mathcal{S}$ and $r = 0, 1$,

$$T_{n,r,s}^{1,\mathcal{T}} = \max_{(\tau,\tau')\in\mathcal{T}^2:\tau\neq\tau'} \left\| \sqrt{n} \left(F_{n,Y_{i0}|T_i=\tau,S_i=s,R_i=r} - F_{n,Y_{i0}|T_i=\tau',S_i=s,R_i=r} \right) \right\|.$$

The randomization procedure proposed in the paper using the transformations $\mathcal{G}_0^{1,\mathcal{S}}$ can be used to obtain p-values for $T_n^{1,\mathcal{S},\mathcal{T}}$ under $H_0^{1,\mathcal{S},\mathcal{T}}$.

Let $(\tau, r) \in \mathcal{T} \times \mathcal{R}$. Let (τ_j, r_j) denote the j^{th} element of $\mathcal{T} \times \mathcal{R}$, then the null hypothesis implied by Proposition 9(b.ii) is given by the following:

$$H_0^{2,\mathcal{S},\mathcal{T}} : F_{Y_{i0}|T_i=\tau_j,S_i=s,R_i=r_j} = F_{Y_{i0}|T_i=\tau_{j+1},S_i=s,R_i=r_{j+1}} \text{ for } j = 1,\ldots,|\mathcal{T}\times\mathcal{R}|-1,\ s \in \mathcal{S}. \tag{3G.6}$$

the test statistic for the above *joint* hypothesis is given by

$$T_{n,m}^{2,\mathcal{S},\mathcal{T}} = \max_{s\in\mathcal{S}} \max_{j=1,\ldots,|\mathcal{T}\times\mathcal{R}|-1} \left\| \sqrt{n} \left(F_{n,Y_{i0}|T_i=\tau_j,S_i=s,R_i=r_j} - F_{n,Y_{i0}|T_i=\tau_{j+1},S_i=s,R_i=r_{j+1}} \right) \right\|,$$

The randomization procedure proposed in the paper using the transformations $\mathcal{G}_0^{2,\mathcal{S}}$ can be used to obtain p-values for the above statistic under $H_0^{2,\mathcal{S},\mathcal{T}}$.

3H Extended Simulations for the Distributional Tests

3H.1 Comparing Different Statistics of the Distributional Hypotheses

In this section, we examine the finite-sample performance of a wider variety of the distributional tests of the IV-R and IV-P assumptions provided in Section 3.4 of the paper. We specifically consider the Kolmogorov-Smirnov (KS) and Cramer-von-Mises (CM) statistics of the simple and joint hypotheses. For the joint hypotheses, we include the probability weighted statistic in addition to the version used in the paper.

For the IV-R assumption, consider the following hypotheses implied by Proposition 5(b.ii) in the paper

$$\begin{aligned} H_0^{1,1}: &\quad Y_{i0}|T_i=1,R_i=0 \overset{d}{=} Y_{i0}|T_i=0,R_i=0, && (CA-TA) \\ H_0^{1,2}: &\quad Y_{i0}|T_i=1,R_i=1 \overset{d}{=} Y_{i0}|T_i=0,R_i=1, && (CR-TR) \\ H_0^{1} &: \quad H_0^{1,1} \ \& \ H_0^{1,2}. && (Joint) \end{aligned} \tag{3H.1}$$

For $r=0,1$, the KS and CM statistics to test $H_0^{1,r+1}$ is given by

$$\begin{aligned} KS_{n,r}^1 &= \max_{i:R_i=r} \left| \sqrt{n} \left(F_{n,Y_{i0}}(y_{i0}|T_i=1,R_i=r) - F_{n,Y_{i0}}(y_{i0}|T_i=0,R_i=r) \right) \right|. \\ CM_{n,r}^1 &= \frac{\sum_{i:R_i=r} \left(\sqrt{n} (F_{n,Y_{i0}}(y_{i0}|T_i=1,R_i=r) - F_{n,Y_{i0}}(y_{i0}|T_i=0,R_i=r) \right)^2}{\sum_{i=1}^n 1\{R_i=r\}} \end{aligned} \tag{3H.2}$$

For the joint hypothesis H_0^1, which is the sharp testable restriction in Proposition 5(b.ii) in the paper, we consider either $KS_{n,m}^1 = \max\{KS_{n,0}^1, KS_{n,1}^1\}$ or $KS_{n,p}^1 = p_{n,0}KS_{n,0}^1 + p_{n,1}KS_{n,1}^1$, where $p_{n,r} = \sum_{i=1}^n 1\{R_i = r\}/n$ for $r=0,1$. $CM_{n,m}^1$ and $CM_{n,p}^1$ are similarly defined.

Table 3D presents the simulation rejection probabilities of the aforementioned statistics of the IV-R assumption. For each simulation design and attrition rate, we report the rejection probabilities for the KS statistics of the simple hypotheses, $KS_{n,0}^1$ and $KS_{n,1}^1$, using asymptotic critical values (KS $(Asym.)$) as a benchmark for the KS (KS (R)) and the CM (CM (R)) statistics using the p-values obtained from the proposed randomization procedure to test H_0^1 ($B = 199$). The different variants of the KS and CM test statistics control size under Designs II and III, where IV-R holds. They also have non-trivial power in finite samples in Designs I and IV, when IV-R is violated. The simulation results for the distributional statistics also illustrate the potential power gains in finite samples from using the attritor subgroup in testing the IV-R assumption. In testing the joint null hypothesis, we find that $KS_{n,m}^1$ and $CM_{n,m}^1$ ($Joint$ (m)) exhibit better finite-sample power properties than $KS_{n,p}^1$ and $CM_{n,p}^1$ ($Joint$ (p)). We also note that the randomization procedure yields rejection probabilities for the two-sample KS statistics, $KS_{n,0}^1$ and $KS_{n,1}^1$, that are very similar to those obtained from the asymptotic critical values. In addition, in our simulation design, the CM statistics generally have better finite-sample power properties than their respective KS statistics, while maintaining comparable size control.

We then examine the finite-sample performance of the distributional statistics of the IV-P assumption. Proposition 5(b.ii) in the paper implies the three simple null hypotheses as well as their joint hypothesis

below,

$$
\begin{aligned}
H_0^{2,1}: &\quad Y_{i0}|T_i=0,R_i=0 \overset{d}{=} Y_{i0}|T_i=0,R_i=1, && (CA-CR)\\
H_0^{2,2}: &\quad Y_{i0}|T_i=0,R_i=1 \overset{d}{=} Y_{i0}|T_i=1,R_i=0, && (CR-TA)\\
H_0^{2,3}: &\quad Y_{i0}|T_i=1,R_i=0 \overset{d}{=} Y_{i0}|T_i=1,R_i=1, && (TA-TR)\\
H_0^{2}: &\quad H_0^{2,1} \;\&\; H_0^{2,2} \;\&\; H_0^{2,3}. && (Joint)
\end{aligned} \tag{3H.3}
$$

Let (τ_j, r_j) denote the j^{th} element of $\mathcal{T}\times\mathcal{R} = \{(0,0),(0,1),(1,0),(1,1)\}$. We can define the KS and CM statistics for $H_0^{2,j}$ for each $j=1,2,3$ by the following,

$$
\begin{aligned}
KS_{n,j}^2 &= \max_{i:(T_i,R_i)\in\{(\tau_j,r_j),(\tau_{j+1},r_{j+1})\}} \left|\sqrt{n}\left(F_{n,Y_{i0}|T_i=\tau_{j-1},R_i=r_{j-1}} - F_{n,Y_{i0}|T_i=\tau_j,R_i=r_j}\right)\right|,\\
CM_{n,j}^2 &= \frac{\sum_{i:(T_i,R_i)\in\{(\tau_j,r_j),(\tau_{j+1},r_{j+1})} \left(\sqrt{n}\left(F_{n,Y_{i0}|T_i=\tau_{j-1},R_i=r_{j-1}} - F_{n,Y_{i0}|T_i=\tau_j,R_i=r_j}\right)\right)^2}{\sum_{i=1}^n 1\left\{(T_i,R_i)\in\{(\tau_j,r_j),(\tau_{j+1},r_{j+1})\}\right\}},
\end{aligned} \tag{3H.4}
$$

The joint hypothesis H_0^2 is tested using the joint statistics $KS_{n,m}^2 = \max_{j=1,2,3} KS_{n,j}^2$ and $CM_{n,m}^2 = \max_{j=1,2,3} CM_{n,j}^2$.

In Table 3E, we report the simulation rejection probabilities for distributional tests of the IV-P assumption. In addition to the aforementioned statistics whose p-values are obtained using the proposed randomization procedure to test H_0^2 ($B = 199$), the table also reports the simulation results for the KS statistics of the simple hypotheses using the asymptotic critical values. Under Designs I, II and IV, IV-P is violated, the rejection probabilities for all the test statistics we consider tend to be higher than the nominal level, as we would expect. The joint KS and CM test statistics behave similarly in this design and have comparable finite-sample power properties to the test statistic of the simple hypothesis (TA-TR), which has the best finite-sample power properties in our simulation design. Finally, in Design III, where IV-P holds, our simulation results illustrate that the test statistics we consider control size.

3H.2 Additional Variants of the Simulation Designs

To illustrate the relative power properties of using the simple vs joint tests of internal validity, we present additional results using variants of the simulation designs. We show the results of the KS tests for the case where $P(R_i = 0|T_i = 0) = 0.15$.[58] For the joint hypotheses, we report the simulation results for the KS statistic that takes the maximum over the individual statistics.

[58] We use an attrition rate of 15% in the control group as reference since that is the average attrition rate in our review of field experiments. See Section 3.2 in the paper for details.

Panel A in Figure 3A displays the simulation rejection probabilities of the tests of the IV-R assumption while Panel B displays the simulation rejection probabilities of the tests of the IV-P assumption. We present these rejection probabilities for alternative parameter values of the designs we consider in Section 3.4 in the paper. *Design II to I* depicts the case in which we vary the proportion of treatment-only responders, p_{01}, from zero to $0.9 \times P(R_i = 0|T_i = 0)$, where $p_{01} = 0$ corresponds to Design II and $p_{01} > 0$ to variants of Design I. *Design III to I* depicts the case in which we vary the correlation parameter between the unobservables in the outcome equation and the unobservables in the response equation, ρ, from zero to one. Hence, $\rho = 0$ corresponds to Design III while $\rho > 0$ corresponds to different versions of Design I. Finally, the results under *Design II to IV* are obtained by fixing $p_{01} = p_{10}$ and varying them from zero to $0.9 \times P(R_i = 0|Ti = 0)$. Design II corresponds to the case in which $p_{01} = p_{10} = 0$ and $p_{01} = p_{10} > 0$ corresponds to different versions of Design IV.

Overall, the simulation results illustrate that the *joint* tests that we propose in Section 3A in the paper have better finite-sample power properties relative to the statistics of the simple null hypotheses. Most notably, the results under *Design II to I* in Panel A of Figure 3A show that when IV-R does not hold (i.e. $p_{01} > 0$), the simulation rejection probabilities of the joint test are generally above the simulation rejection probabilities of the simple test that only uses the respondents.

Table 3D: Simulation Results on the KS & CM Randomization Test of IV-R

Design	Att. Rate		KS (*Asym.*)		KS (R)				CM(R)			
	C	T	CR-TR	CA-TA	CR-TR	CA-TA	Joint (m)	Joint (p)	CR-TR	CA-TA	Joint (m)	Joint (p)
Differential Attrition Rates + Monotonicity + $(U_{i0}, U_{i1}) \not\perp (R_i(0), R_i(1))$												
I	0.050	0.025	0.058	0.316	0.058	0.324	0.324	0.081	0.058	0.353	0.353	0.285
	0.100	0.050	0.066	0.589	0.071	0.582	0.582	0.157	0.072	0.636	0.636	0.568
	0.150	0.100	0.067	0.460	0.067	0.483	0.483	0.167	0.069	0.544	0.544	0.460
	0.200	0.150	0.070	0.392	0.073	0.412	0.412	0.180	0.069	0.462	0.462	0.385
	0.300	0.200	0.111	0.790	0.123	0.801	0.801	0.502	0.135	0.855	0.855	0.803
Equal Attrition Rates + Monotonicity + $(U_{i0}, U_{i1}) \not\perp (R_i(0), R_i(1))^{\dagger}$												
II	0.050	0.050	0.052	0.059	0.053	0.062	0.062	0.052	0.054	0.056	0.056	0.061
	0.100	0.100	0.049	0.054	0.053	0.056	0.056	0.050	0.054	0.054	0.054	0.053
	0.150	0.150	0.044	0.049	0.049	0.055	0.055	0.051	0.049	0.054	0.054	0.055
	0.200	0.200	0.052	0.044	0.052	0.050	0.050	0.058	0.052	0.049	0.049	0.052
	0.300	0.300	0.051	0.043	0.051	0.042	0.043	0.053	0.049	0.047	0.048	0.057
Differential Attrition Rates + Monotonicity + $(U_{i0}, U_{i1}) \perp (R_i(0), R_i(1))$ (*Example 1*)*												
III	0.050	0.025	0.049	0.051	0.054	0.052	0.052	0.056	0.048	0.051	0.051	0.049
	0.100	0.050	0.047	0.042	0.050	0.046	0.046	0.047	0.053	0.047	0.047	0.043
	0.150	0.100	0.047	0.038	0.052	0.045	0.045	0.047	0.049	0.049	0.049	0.048
	0.200	0.150	0.054	0.031	0.053	0.036	0.036	0.047	0.055	0.036	0.036	0.044
	0.300	0.200	0.050	0.043	0.050	0.043	0.043	0.050	0.051	0.042	0.042	0.050
Equal Attrition Rates + Violation of Monotonicity + $(U_{i0}, U_{i1}) \not\perp (R_i(0), R_i(1))$ (*Example 2*)												
IV	0.050	0.050	0.059	0.332	0.065	0.329	0.329	0.093	0.067	0.375	0.375	0.302
	0.100	0.100	0.102	0.569	0.102	0.577	0.577	0.230	0.116	0.663	0.663	0.593
	0.150	0.150	0.178	0.740	0.190	0.758	0.758	0.465	0.211	0.816	0.816	0.805
	0.200	0.200	0.313	0.854	0.319	0.859	0.859	0.709	0.368	0.917	0.916	0.910
	0.300	0.300	0.683	0.970	0.680	0.972	0.974	0.974	0.760	0.985	0.991	0.996

Notes: The above table presents the rejection probabilities of the KS and CM tests for the simple and joint null hypotheses in (3H.1). We use the nominal level $\alpha = 0.05$, 2,000 simulation replications and $n = 2,000$. C denotes the control group, T denotes the treatment group. $KS(Asym.)$ refers to the two-sample KS test using the asymptotic critical values. $KS(R)$ and $CM(R)$ refer to the randomization KS and CM tests, respectively, for the simple and joint hypotheses. *Joint* (m) and *Joint* (p) denote the randomization procedure applied to $KS^1_{n,m}$ ($CM^1_{n,m}$) and $KS^1_{n,p}$ ($CM^1_{n,p}$), respectively. Additional details of the design are provided in Table 3.4 in the paper.
† (*) indicates IV-R only (IV-P).

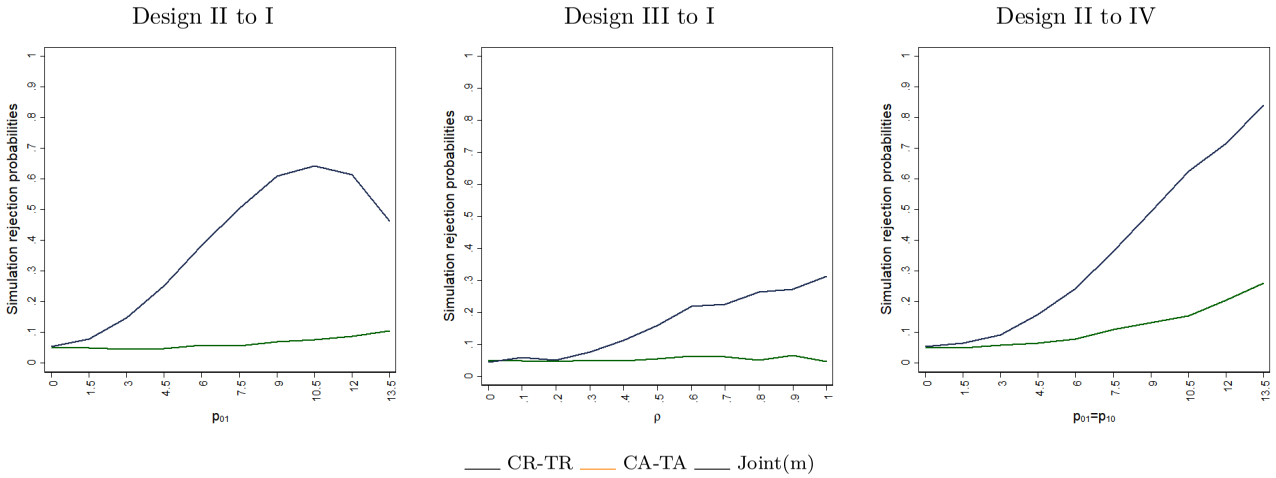
Table 3E: Simulation Results on the KS & CM Randomization Test of IV-P

Design	Att. Rate		KS (*Asym.*)			KS (R)				CM(R)			
	C	T	CA-CR	CR-TA	TA-TR	CA-CR	CR-TA	TA-TR	*Joint* (m)	CA-CR	CR-TA	TA-TR	*Joint* (m)
Differential Attrition Rates + Monotonicity + $(U_{i0}, U_{i1}) \not\perp (R_i(0), R_i(1))$													
I	0.050	0.025	0.051	0.451	0.456	0.064	0.482	0.485	0.476	0.053	0.492	0.497	0.483
	0.100	0.050	0.053	0.746	0.787	0.055	0.763	0.801	0.787	0.058	0.806	0.837	0.824
	0.150	0.100	0.414	0.970	0.980	0.420	0.969	0.978	0.980	0.463	0.983	0.986	0.989
	0.200	0.150	0.865	0.999	0.998	0.870	0.998	0.998	1.000	0.902	1.000	0.999	1.000
	0.300	0.200	0.774	1.000	1.000	0.771	1.000	1.000	1.000	0.825	1.000	1.000	1.000
Equal Attrition Rates + Monotonicity + $(U_{i0}, U_{i1}) \not\perp (R_i(0), R_i(1))^{\dagger}$													
II	0.050	0.050	0.772	0.788	0.788	0.780	0.797	0.804	0.902	0.831	0.840	0.841	0.939
	0.100	0.100	0.984	0.983	0.980	0.985	0.981	0.981	0.999	0.994	0.989	0.986	1.000
	0.150	0.150	1.000	1.000	0.998	1.000	1.000	0.998	1.000	1.000	1.000	0.999	1.000
	0.200	0.200	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	0.300	0.300	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Differential Attrition Rates + Monotonicity + $(U_{i0}, U_{i1}) \perp (R_i(0), R_i(1))$ (*Example 1*)*													
III	0.050	0.025	0.040	0.042	0.043	0.044	0.050	0.051	0.050	0.047	0.053	0.053	0.054
	0.100	0.050	0.051	0.041	0.048	0.058	0.052	0.052	0.055	0.056	0.050	0.057	0.056
	0.150	0.100	0.040	0.051	0.052	0.046	0.056	0.057	0.059	0.047	0.054	0.055	0.059
	0.200	0.150	0.037	0.040	0.045	0.041	0.046	0.050	0.048	0.046	0.045	0.054	0.050
	0.300	0.200	0.048	0.044	0.044	0.050	0.049	0.046	0.048	0.049	0.044	0.051	0.054
Equal Attrition Rates + Violation of Monotonicity + $(U_{i0}, U_{i1}) \not\perp (R_i(0), R_i(1))$ (*Example 2*)													
IV	0.050	0.050	0.075	0.325	0.361	0.082	0.350	0.384	0.311	0.097	0.363	0.407	0.342
	0.100	0.100	0.113	0.548	0.668	0.125	0.558	0.681	0.582	0.152	0.605	0.742	0.661
	0.150	0.150	0.169	0.683	0.854	0.180	0.694	0.858	0.792	0.220	0.756	0.908	0.861
	0.200	0.200	0.234	0.759	0.947	0.239	0.762	0.950	0.913	0.288	0.822	0.974	0.952
	0.300	0.300	0.371	0.805	0.999	0.376	0.813	0.999	0.998	0.440	0.875	1.000	1.000

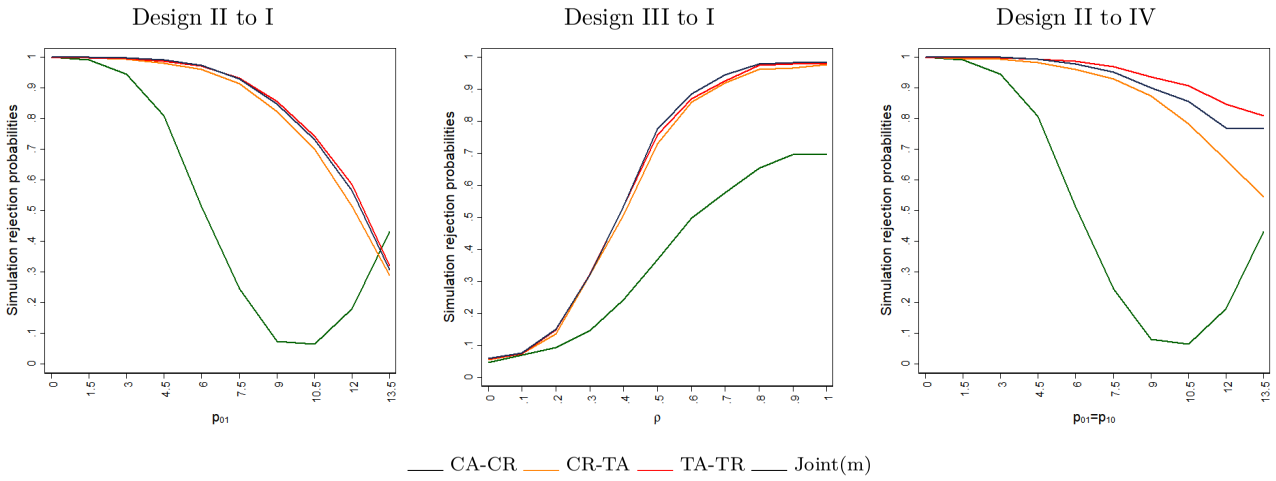
Notes: The above table presents the rejection probabilities of the KS and CM tests for the simple and joint null hypotheses in (3H.3). We use the nominal level $\alpha = 0.05$, 2,000 simulation replications and $n = 2,000$. C denotes the control group, T denotes the treatment group. $KS(Asym.)$ refers to the two-sample test using the asymptotic critical values. $KS(R)$ and $CM(R)$ refer to the randomization KS and CM tests, respectively, for the simple and joint hypotheses. *Joint* (m) denotes the randomization procedure applied to $KS^2_{n,m}$ ($CM^2_{n,m}$). Additional details of the design are provided in Table 3.4 in the paper.
† (*) indicates IV-R only (IV-P).

Figure 3A: Additional Simulation Analysis for the KS Statistic of Internal Validity

Panel A. Internal Validity for Respondents



Panel B. Internal Validity for the Study Population



3I List of Papers Included in the Review of Field Experiments

Abeberese, Ama Baafra, Todd J. Kumler, and Leigh L. Linden. 2014. "Improving Reading Skills by Encouraging Children to Read in School: A Randomized Evaluation of the Sa Aklat Sisikat Reading Program in the Philippines." *Journal of Human Resources*, 49 (3): 611–33.

Abdulkadiroğlu, A., Angrist, J. D., Dynarski, S. M., Kane, T. J., & Pathak, P. A. (2011). Accountability and Flexibility in Public Schools: Evidence from Boston's Charters And Pilots. *Quarterly Journal of Economics*, 126(2), 699-748.

Aker, J. C., Ksoll, C., & Lybbert, T. J. (2012). Can Mobile Phones Improve Learning? Evidence from a Field Experiment in Niger. *American Economic Journal: Applied Economics*, 4(4), 94-120.

Ambler, K. (2015). Don't tell on me: Experimental evidence of asymmetric information in transnational households. *Journal of Development Economics*, 113, 52-69.

Ambler, K., Aycinena, D., & Yang, D. (2015). Channeling Remittances to Education: A Field Experiment among Migrants from El Salvador. *American Economic Journal: Applied Economics*, 7(2), 207-232.

Anderson, E. T., & Simester, D. I. (2010). Price Stickiness and Customer Antagonism. *Quarterly Journal of Economics*, 125(2), 729–765.

Ashraf, N., Aycinena, D., Martínez A., C., & Yang, D. (2015). Savings in Transnational Households: A Field Experiment among Migrants from El Salvador. *Review of Economics and Statistics*, 97(2), 332-351.

Ashraf, N., Berry, J., & Shapiro, J. M. (2010). Can Higher Prices Stimulate Product Use? Evidence from a Field Experiment in Zambia. *American Economic Review*, 100(5), 2383-2413.

Attanasio, O., Augsburg, B., De Haas, R., Fitzsimons, E., & Harmgart, H. (2015). The Impacts of Microfinance: Evidence from Joint-Liability Lending in Mongolia. *American Economic Journal: Applied Economics*, 7(1), 90-122.

Augsburg, B., De Haas, R., Harmgart, H., & Meghir, C. (2015). The Impacts of Microcredit: Evidence from Bosnia and Herzegovina. *American Economic Journal: Applied Economics*, 7(1), 183-203. Avitabile, Ciro. 2012. "Does Information Improve the Health Behavior of Adults Targeted by a Conditional Transfer Program?" *Journal of Human Resources*, 47 (3): 785–825.

Avvisati, F., Gurgand, M., Guyon, N., & Maurin, E. (2014). Getting Parents Involved: A Field Experiment in Deprived Schools. *Review of Economic Studies*, 81(1), 57-83.

Baird, S., McIntosh, C., & Özler, B. (2011). Cash or Condition? Evidence from a Cash Transfer Experiment. *Quarterly Journal of Economics*, 126(4), 1709-1753.

Barham, T. (2011). A healthier start: The effect of conditional cash transfers on neonatal and infant

mortality in rural Mexico. *Journal of Development Economics*, 94(1), 74-85.

Barton, J., Castillo, M., & Petrie, R. (2014). What Persuades Voters? A Field Experiment on Political Campaigning. *Economic Journal*, 124(574), F293–F326.

Basu, K., & Wong, M. (2015). Evaluating seasonal food storage and credit programs in east Indonesia. *Journal of Development Economics*, 115, 200-216.

Bauchet, J., Morduch, J., & Ravi, S. (2015). Failure vs. displacement: Why an innovative anti-poverty program showed no net impact in South India. *Journal of Development Economics*, 116, 1-16.

Bengtsson, N., & Engström, P. (2014). Replacing Trust with Control: A Field Test of Motivation Crowd Out Theory. *Economic Journal*, 124(577), 833-858.

Berry, James. 2015. "Child Control in Education Decisions: An Evaluation of Targeted Incentives to Learn in India." *Journal of Human Resources* 50 (4): 1051–80.

Bettinger, E. P. (2012). Paying to Learn: The Effect of Financial Incentives on Elementary School Test Scores. *Review of Economics and Statistics*, 94(3), 686-698.

Beuermann, D. W., Cristia, J., Cueto, S., Malamud, O., & Cruz-Aguayo, Y. (2015). One Laptop per Child at Home: Short-Term Impacts from a Randomized Experiment in Peru. *American Economic Journal: Applied Economics*, 7(2), 53-80.

Bianchi, M., & Bobba, M. (2013). Liquidity, Risk, and Occupational Choices. *Review of Economic Studies*, 80(2), 491-511.

Björkman, M., & Svensson, J. (2009). Power to the People: Evidence from a Randomized Field Experiment on Community-Based Monitoring in Uganda. *Quarterly Journal of Economics*, 124(2), 735-769.

Blattman, C., Fiala, N., & Martinez, S. (2014). Generating Skilled Self-Employment in Developing Countries: Experimental Evidence from Uganda. *Quarterly Journal of Economics*, 129(2), 697-752.

Bloom, N., Eifert, B., Mahajan, A., McKenzie, D., & Roberts, J. (2013). Does Management Matter? Evidence from India. *Quarterly Journal of Economics*, 128(1), 1-51.

Bloom, N., Liang, J., Roberts, J., & Ying, Z. J. (2015). Does Working from Home Work? Evidence from a Chinese Experiment. *Quarterly Journal of Economics*, 130(1), 165-218.

Bobonis, G. J., & Finan, F. (2009). Neighborhood Peer Effects in Secondary School Enrollment Decisions. *Review of Economics and Statistics*, 91(4), 695-716.

Bruhn, M., Ibarra, G. L., & McKenzie, D. (2014). The minimal impact of a large-scale financial education program in Mexico City. *Journal of Development Economics*, 108, 184-189.

Bryan, G., Chowdhury, S., & Mobarak, A. M. (2014). Underinvestment in a Profitable Technology: The Case of Seasonal Migration in Bangladesh. *Econometrica*, 82(5), 1671-1748.

Cai, H., Chen, Y., Fang, H., & Zhou, L.-A. (2015). The Effect of Microinsurance on Economic Activities: Evidence from a Randomized Field Experiment. Review of Economics and Statistics.

Charness, G., & Gneezy, U. (2009). Incentives to Exercise. *Econometrica*, 77(3), 909-931.

Chetty, R., & Saez, E. (2013). Teaching the Tax Code: Earnings Responses to an Experiment with EITC Recipients. *American Economic Journal: Applied Economics*, 5(1), 1-31.

Collier, P., & Vicente, P. C. (2014). Votes and Violence: Evidence from a Field Experiment in Nigeria. *Economic Journal*, 124(574), F327–F355.

Crépon, B., Devoto, F., Duflo, E., & Parienté, W. (2015). Estimating the Impact of Microcredit on Those Who Take It Up: Evidence from a Randomized Experiment in Morocco. *American Economic Journal: Applied Economics*, 7(1), 123-150.

Cunha, J. M. (2014). Testing Paternalism: Cash versus In-Kind Transfers. *American Economic Journal: Applied Economics*, 6(2), 195-230.

De Grip, A., & Sauermann, J. (2012). The Effects of Training on Own and Co-worker Productivity: Evidence from a Field Experiment. *Economic Journal*, 122(560), 376-399.

de Mel, S., McKenzie, D., & Woodruff, C. (2014). Business training and female enterprise start-up, growth, and dynamics: Experimental evidence from Sri Lanka. *Journal of Development Economics*, 106, 199-210.

De Mel, S., McKenzie, D., & Woodruff, C. (2012). Enterprise Recovery Following Natural Disasters. *Economic Journal*, 122(559), 64-91.

de Mel, S., McKenzie, D., & Woodruff, C. (2013). The Demand for, and Consequences of, Formalization among Informal Firms in Sri Lanka. *American Economic Journal: Applied Economics*, 5(2), 122-150.

Dinkelman, T., & Martínez A., C. (2014). Investing in Schooling In Chile: The Role of Information about Financial Aid for Higher Education. *Review of Economics and Statistics*, 96(2), 244-257.

Doi, Y., McKenzie, D., & Zia, B. (2014). Who you train matters: Identifying combined effects of financial education on migrant households. *Journal of Development Economics*, 109, 39–55.

Duflo, E., Dupas, P., & Kremer, M. (2011). Peer Effects, Teacher Incentives, and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya. *American Economic Review*, 101(5), 1739-1774.

Duflo, E., Greenstone, M., Pande, R., & Ryan, N. (2013). Truth-telling by Third-party Auditors and the Response of Polluting Firms: Experimental Evidence from India. *Quarterly Journal of Economics*, 128(4), 1499-1545.

Duflo, E., Hanna, R., & Ryan, S. P. (2012). Incentives Work: Getting Teachers to Come to School. *American Economic Review*, 102(4), 1241-1278.

Dupas, P., & Robinson, J. (2013). Savings Constraints and Microenterprise Development: Evidence from

a Field Experiment in Kenya. *American Economic Journal: Applied Economics*, 5(1), 163-192.

Edmonds, E. V, & Shrestha, M. (2014). You get what you pay for: Schooling incentives and child labor. *Journal of Development Economics*, 111, 196-211.

Fafchamps, M., McKenzie, D., Quinn, S., & Woodruff, C. (2014). Microenterprise growth and the flypaper effect: Evidence from a randomized experiment in Ghana. *Journal of Development Economics*, 106(Supplement C), 211-226.

Fafchamps, M., & Vicente, P. C. (2013). Political violence and social networks: Experimental evidence from a Nigerian election. *Journal of Development Economics*, 101(Supplement C), 27-48.

Ferraro, P. J., & Price, M. K. (2013). Using Nonpecuniary Strategies to Influence Behavior: Evidence from a Large-Scale Field Experiment. *Review of Economics and Statistics*, 95(1), 64-73.

Finkelstein, A., Taubman, S., Wright, B., Bernstein, M., Gruber, J., Newhouse, J. P., Baicker, K. (2012). The Oregon Health Insurance Experiment: Evidence from the First Year. *Quarterly Journal of Economics*, 127(3), 1057-1106.

Fryer, J. R. G. (2011). Financial Incentives and Student Achievement: Evidence from Randomized Trials. *Quarterly Journal of Economics*, 126(4), 1755-1798.

Fryer, J. R. G. (2014). Injecting Charter School Best Practices into Traditional Public Schools: Evidence from Field Experiments. *Quarterly Journal of Economics*, 129(3), 1355-1407.

Gertler, P. J., Martinez, S. W., & Rubio-Codina, M. (2012). Investing Cash Transfers to Raise Long-Term Living Standards. *American Economic Journal: Applied Economics*, 4(1), 164-192.

Giné, X., Goldberg, J., & Yang, D. (2012). Credit Market Consequences of Improved Personal Identification: Field Experimental Evidence from Malawi. *American Economic Review*, 102(6), 2923-2954.

Giné, X., & Karlan, D. S. (2014). Group versus individual liability: Short and long term evidence from Philippine microcredit lending groups. *Journal of Development Economics*, 107, 65-83.

Hainmueller, J., Hiscox, M. J., & Sequeira, S. (2015). Consumer Demand for Fair Trade: Evidence from a Multistore Field Experiment. *Review of Economics and Statistics*, 97(2), 242-256.

Hanna, R., Mullainathan, S., & Schwartzstein, J. (2014). Learning Through Noticing: Theory and Evidence from a Field Experiment. *Quarterly Journal of Economics*, 129(3), 1311-1353.

Hidrobo, M., Hoddinott, J., Peterman, A., Margolies, A., & Moreira, V. (2014). Cash, food, or vouchers? Evidence from a randomized experiment in northern Ecuador. *Journal of Development Economics*, 107, 144-156.

Jackson, C. K., & Schneider, H. S. (2015). Checklists and Worker Behavior: A Field Experiment. *American Economic Journal: Applied Economics*, 7(4), 136-168.

Jacob, B. A., Kapustin, M., & Ludwig, J. (2015). The Impact of Housing Assistance on Child Outcomes:

Evidence from a Randomized Housing Lottery. *Quarterly Journal of Economics*, 130(1), 465-506.

Jensen, R. (2012). Do Labor Market Opportunities Affect Young Women's Work and Family Decisions? Experimental Evidence from India. *Quarterly Journal of Economics*, 127(2), 753-792.

Jensen, R. T., & Miller, N. H. (2011). Do Consumer Price Subsidies Really Improve Nutrition? *Review of Economics and Statistics*, 93(4), 1205-1223.

Just, David R., and Joseph Price. 2013. "Using Incentives to Encourage Healthy Eating in Children." *Journal of Human Resources* 48 (4): 855–72.

Karlan, D., Osei, R., Osei-Akoto, I., & Udry, C. (2014). Agricultural Decisions after Relaxing Credit and Risk Constraints. *Quarterly Journal of Economics*, 129(2), 597-652.

Karlan, D., & Valdivia, M. (2011). Teaching Entrepreneurship: Impact of Business Training on Microfinance Clients and Institutions. *Review of Economics and Statistics*, 93(2), 510-527.

Kazianga, H., de Walque, D., & Alderman, H. (2014). School feeding programs, intrahousehold allocation and the nutrition of siblings: Evidence from a randomized trial in rural Burkina Faso. *Journal of Development Economics*, 106, 15-34.

Kendall, C., Nannicini, T., & Trebbi, F. (2015). How Do Voters Respond to Information? Evidence from a Randomized Campaign. *American Economic Review*, 105(1), 322-353.

Kling, J. R., Mullainathan, S., Shafir, E., Vermeulen, L. C., & Wrobel, M. V. (2012). Comparison Friction: Experimental Evidence from Medicare Drug Plans. *Quarterly Journal of Economics*, 127(1), 199-235.

Kremer, M., Leino, J., Miguel, E., & Zwane, A. P. (2011). Spring Cleaning: Rural Water Impacts, Valuation, and Property Rights Institutions. *Quarterly Journal of Economics*, 126(1), 145-205.

Labonne, J. (2013). The local electoral impacts of conditional cash transfers: Evidence from a field experiment. *Journal of Development Economics*, 104, 73–88.

Lalive, R., & Cattaneo, M. A. (2009). Social Interactions and Schooling Decisions. *Review of Economics and Statistics*, 91(3), 457-477.

Macours, K., Schady, N., & Vakis, R. (2012). Cash Transfers, Behavioral Changes, and Cognitive Development in Early Childhood: Evidence from a Randomized Experiment. *American Economic Journal: Applied Economics*, 4(2), 247-273.

Macours, K., & Vakis, R. (2014). Changing Households' Investment Behaviour through Social Interactions with Local Leaders: Evidence from a Randomised Transfer Programme. *Economic Journal*, 124(576), 607-633.

Meredith, J., Robinson, J., Walker, S., & Wydick, B. (2013). Keeping the doctor away: Experimental evidence on investment in preventative health products. *Journal of Development Economics*, 105, 196–210.

Muralidharan, K., & Sundararaman, V. (2011). Teacher Performance Pay: Experimental Evidence from India. Journal of Political Economy, 119(1), 39-77.

Muralidharan, K., & Sundararaman, V. (2015). The Aggregate Effect of School Choice: Evidence from a Two-Stage Experiment in India. *Quarterly Journal of Economics*, 130(3), 1011-1066.

Olken, B. A., Onishi, J., & Wong, S. (2014). Should Aid Reward Performance? Evidence from a Field Experiment on Health and Education in Indonesia. *American Economic Journal: Applied Economics*, 6(4), 1-34.

Pallais, A. (2014). Inefficient Hiring in Entry-Level Labor Markets. *American Economic Review*, 104(11), 3565-3599.

Pomeranz, D. (2015). No Taxation without Information: Deterrence and Self-Enforcement in the Value Added Tax. *American Economic Review*, 105(8), 2539-2569.

Powell-Jackson, T., Hanson, K., Whitty, C. J. M., & Ansah, E. K. (2014). Who benefits from free healthcare? Evidence from a randomized experiment in Ghana. *Journal of Development Economics*, 107, 305-319.

Pradhan, M., Suryadarma, D., Beatty, A., Wong, M., Gaduh, A., Alisjahbana, A., & Artha, R. P. (2014). Improving Educational Quality through Enhancing Community Participation: Results from a Randomized Field Experiment in Indonesia. *American Economic Journal: Applied Economics*, 6(2), 105-126.

Prina, S. (2015). Banking the poor via savings accounts: Evidence from a field experiment. *Journal of Development Economics*, 115, 16-31.

Reichert, Arndt R. 2015. "Obesity, Weight Loss, and Employment Prospects: Evidence from a Randomized Trial." *Journal of Human Resources* 50 (3): 759–810.

Royer, H., Stehr, M., & Sydnor, J. (2015). Incentives, Commitments, and Habit Formation in Exercise: Evidence from a Field Experiment with Workers at a Fortune-500 Company. *American Economic Journal: Applied Economics*, 7(3), 51-84.

Seshan, G., & Yang, D. (2014). Motivating migrants: A field experiment on financial decision-making in transnational households. *Journal of Development Economics*, 108, 119-127.

Stutzer, A., Goette, L., & Zehnder, M. (2011). Active Decisions and Prosocial Behaviour: a Field Experiment on Blood Donation. *Economic Journal*, 121(556), F476-F493.

Szabó, A., & Ujhelyi, G. (2015). Reducing nonpayment for public utilities: Experimental evidence from South Africa. *Journal of Development Economics*, 117, 20–31.

Tarozzi, A., Mahajan, A., Blackburn, B., Kopf, D., Krishnan, L., & Yoong, J. (2014). Micro-loans, insecticide-treated bednets, and malaria: Evidence from a randomized controlled trial in Orissa, India. *American Economic Review*, 104, 1909-41.

Thornton, R. L. (2012). HIV testing, subjective beliefs and economic behavior. *Journal of Development Economics*, 99(2), 300-313.

Valdivia, M. (2015). Business training plus for female entrepreneurship? Short and medium-term experimental evidence from Peru. *Journal of Development Economics*, 113, 33-51.

Vicente, P. C. (2014). Is Vote Buying Effective? Evidence from a Field Experiment in West Africa. *Economic Journal*, 124(574), F356-F387.

Walters, C. R. (2015). Inputs in the Production of Early Childhood Human Capital: Evidence from Head Start. *American Economic Journal: Applied Economics*, 7(4), 76-102.

Wilson, N. L., Xiong, W., & Mattson, C. L. (2014). Is sex like driving? HIV prevention and risk compensation. *Journal of Development Economics*, 106, 78-91.

ACKNOWLEDGMENTS

We thank Alberto Abadie, Josh Angrist, Stephen Boucher, Federico Bugni, Pamela Jakiela, Tae-hwy Lee, Jia Li, Aprajit Mahajan, Matthew Masten, Craig McIntosh, David McKenzie, Adam Rosen, Monica Singhal and Aman Ullah for helpful discussions.