I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 12094

The Impact of Migration on Family Left Behind: Estimation in Presence of Intra-Household Selection of Migrants

Elie Murard

JANUARY 2019

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 12094

The Impact of Migration on Family Left Behind: Estimation in Presence of Intra-Household Selection of Migrants

Elie Murard
IZA and Paris School of Economics

JANUARY 2019

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

The Impact of Migration on Family Left Behind: Estimation in Presence of Intra-Household Selection of Migrants

This paper reexamines the literature on the impact of migration on household members left behind at origin. The empirical problem previous studies address is the self-selection of households into migration, i.e. the endogenous decision as to whether or not send a migrant. Yet, the subsequent selection of which family members migrate and which stay behind generates additional identification problems that have remained largely ignored. To tackle this second form of selectivity *within the households*, I model the behavior of families using latent stratification and potential outcome (Imbens and Angrist, 1994; Rubin, 1974). I show that the point-identification of the causal impact of migration requires strong behavioral assumptions rarely satisfied even with ideal experimental data. As a practical solution, I derive non parametric bounds under different sets of weaker assumptions. Using Mexican panel data, I show that standard estimates ignoring the intra-household selection into migration may suffer from substantial bias.

JEL Classification: C21, F22, J61, O15

Keywords: sample selection, migration, selectivity, bounds, principal stratification

Corresponding author:
Elie Murard
IZA
Schaumburg-Lippe-Str. 5-9
53113 Bonn
Germany

E-mail: murard@iza.org

1 Introduction

With more than 3% of the world population living outside the country of their birth, the effect of international migration on the source countries of origin has become an urgent policy question. Migration often separates families, with some households members migrating and others staying behind in the country of origin. A growing literature explores how migration affects the various dimensions of the welfare of family members left behind , such as education, health, labor supply, or consumption – see Antman (2013) and Adams (2011) for a comprehensive overview.

The empirical problem this literature confronts is the the self-selection of households into migration. Non-random selection *across* households arises because family decisions as to whether send migrant(s) are endogenous. The factors that cause households to engage in migration also affect the outcome of interest for the researcher, thereby confounding the estimation of the causal effect of migration. [2]

Yet a second form of selection arises as the households also select which family members migrate and which stay behind at origin. The family decision as to who migrate and who stay at origin is likely to be influenced by factors related to the outcome of interest, such as individual labor supply or education for instance. In consequence, the selectivity *within the household* poses an additional threat to the identification of the causal migration impact. This second source of endogeneity has largely been ignored in the existing literature.

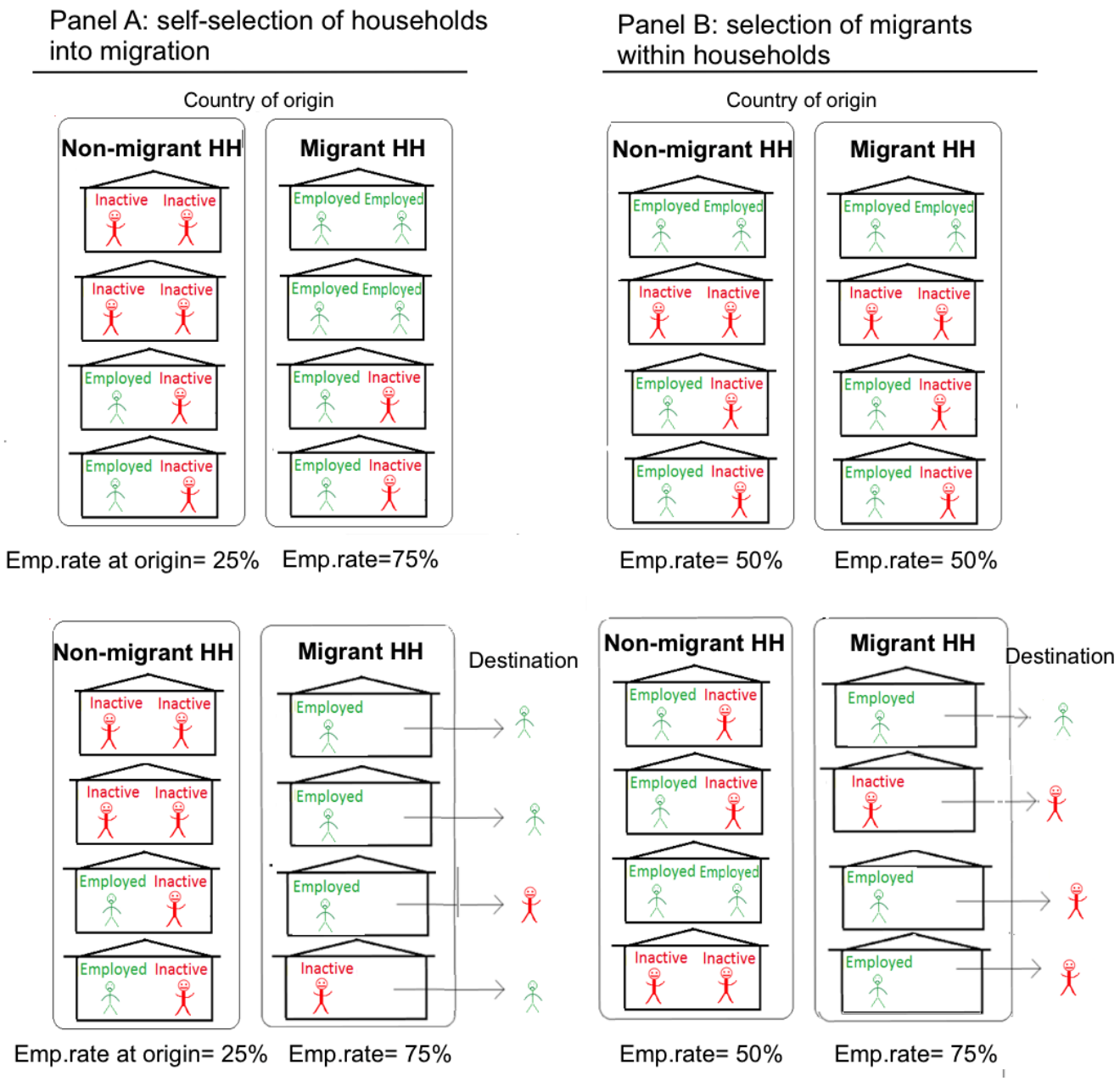
To illustrate the problem, let me consider an economy composed of two-person households of three different types : households in which one member is employed and the other is inactive, households in which both members are employed, and households in which both members are inactive. There are height households in total: four of the first type, two of the second type and two of the third type. So the employment rate in this economy is 50%. I assume that in 50% of the households one member migrates. I also assume that the migration of one household member has no causal effect on the labor force participation of the other member left behind at origin. I then examine the extent to which the estimates an econometrician is likely to produce might be biased by the two different sort of selection, either across or within households.

Panel A in figure 1 illustrates the standard problem of self-selection of households into migration. For example, households where both members are working are more likely to engage in migration because they have higher earnings and can afford the upfront costs of migration. The initial employment rate is, say, 75% among migrant households and 25% among non-migrant households. Assuming away intra-household selection of migrants, the employment rate among left-behinds in migrant households is still 50% higher than among non-migrant households after that the migrants have left. When comparing households with and without migrants, a researcher may therefore conclude that migration increases the labor force participation of non-migrants while the true effect is null.

Even in the absence of selection across households, the intra-household selection of migrants may generate significant bias in the estimates of the impact of migration. To illustrate how, let me consider the following example in Panel B of figure 1 in which I assume that household participation into migration is

[2]The literature uses different empirical strategies to address this identification problem, such as instrumental variables , fixed-effects approaches using panel data, or experimental approaches using randomized or natural experiments (McKenzie and Yang, 2010).

Figure 1: A graphical explanation of selection into migration across and within households in terms of employment status



random, i.e. unrelated to household characteristics. In particular, the initial employment rate is identical between migrant and non-migrant households. However, the selection of which member migrate and which stay behind may not be random. Household members who are inactive may be more likely to migrate than members employed in a regular job at origin (e.g. because of lower opportunity cost). Consequently, after that migrants have left the house, the employment rate among left-behinds in migrant households is on average higher than among non-migrant households. This composition effect biases estimates obtained by comparing employment rate between migrant and non-migrant households, even though they were identical before the departure of the migrant.

This issue have not drawn much empirical attention possibly because that the identification problem that arises is not obvious at first sight. It is not manifest that selection within households is relevant for the estimation of the migration impact across different households, with and without migrants. However, the direct comparison of individuals left behind in migrant households with individuals living in non-migrant households is not satisfactory, neither empirically nor conceptually. This is because, before migration, households at origin are composed of two different type of counterfactual members : potential migrants, i.e. members who would migrate if the family engages in migration, and never-migrants, i.e. members who stay behind in any case. Since migrants are likely to be selected within the family, Gibson et al. (2011) underlines that *the appropriate comparison group for the remaining individuals in migrant households are the group of individuals who would stay behind in non-migrant households* in the case these households were to send a migrant. The identification problem is that, in household without migrants, the econometrician does not observe which member is the potential migrant and which members would stay behind. [3]

The question of intra-household selection into migration has received quite little interest to date, with only two recent studies providing some evidence. Using a migration lottery in Tonga (giving the opportunity to emigrate to New Zealand), Gibson et al. (2011) show strong evidence of positive selectivity into migration among the working-age adult living in the same household. They find that individuals who would possibly migrate are more educated and have twice the weekly income at origin of the same age adults who would stay behind [4]. Using a unique survey on a multi-sited and matched sample of Senegalese migrants at destination (France, Italy, Mauritania) and their household of origin in Senegal, Chort and Senne (2013) explore the key determinants driving the intra-household selection of the migrants. They find that households select as migrants not only the members with higher comparative advantages in earnings at destination, but also those with higher remittances potentials, conditional on earnings.

In this paper, I investigate the implications of this form of selectivity *within the household* for the causal estimation of the impact of migration. I contribute to the literature in different ways. First, I use the approach of principal stratification and potential outcomes (Frangakis and Rubin, 2002; Rubin, 1974) to model the identification problem in presence of double-selection into migration (across and within households). I show what assumptions are implicitly made when the second form of intra-household selectivity is ignored and the consequences of a violation of these. In a general framework, I show that the identification of the causal impact migration is generally not possible, even with ideal experimental data. I therefore turn to partial identification as reasonable solution in practice. I derive non parametric bounds on the causal effect of migration under different sets of transparent behavioral assumptions.

Second, using Mexican panel data (MxFLS survey), I revisit the estimates of Murard (2013) of the effects of migration on the labor supply of young women left behind in Mexico. The bounds suggest that estimates ignoring the intra-household selection of migrants may overstate the true magnitude of

[3] In some context where migration is almost exclusively confined to men, it is possible to abstract from the intra-household selection problem by looking into the impact of migration on women since they almost never migrate themselves (Binzel and Assaad, 2011). How owever, in countries like Mexico (or Indonesia) , the individual characteristics of the migrants are very heterogeneous across households. Both males and females migrate, as well as young and older. The distinction between potential migrants and never-migrants– based on observables characteristics – is far from clear-cut.

[4] This difference being significant in a regression controlling for household fixed effect and gender

the impact of migration on the participation in the labor market. However, bounds tend to confirm the findings in Murard (2013) that migration causes a re-allocation of labor away from non-agricultural jobs to self-employed activities. In a second application, I investigate the effect of migration on school attendance among children left behind in Mexico. Unadjusted point estimates indicate significant negative impacts on school attendance for which the bounds do not to reject zero effect . This result suggests that children may be positively selected into migration among households participating in migration.

Another important finding is that standard IV estimators using instruments for household migration are biased even if there is no systematic selection of migrants within households. The bias in the IV estimates arises because migration (induced by the instrument) generates an endogenous sample selection which the IV estimates do not take into account. In case of absence of intra-household selection, I propose an alternative estimator that converges to the causal impact of migration.

This paper is closely related to the recent work of Steinmayr (2014). The author addresses the problem that migration of the entire household creates for the identification of the causal effect of migration on members left behind at origin. I use the same methodological approach – i.e. principal stratification and latent outcome – to model migration decisions within the family.

Steinmayr emphasizes the issue of whole-household migration and proposes various estimators correcting for what he refers to as an endogenous "invisible sample selection". Yet, endogenous sample selection – i.e. the fact that individuals staying at origin are not an random subset of the original sample before migration – may also be driven by selection of migrants within households, which may be empirically more frequent than whole-household migration. [5] More importantly, the behavioral assumptions made by Steinmayr (2014) about the migration decision process are quite restrictive. The key assumption of his econometric framework is that one household member (the child) would not migrate if the other member (the adult) does not migrate as well. Although justified in his context, this framework imposes a hierarchically ordered process where one member is constrained to be the "principal migrant" and the second to be a "tied-mover" who can only accompany the former. In this paper, I propose a more general model by allowing every household member to migrate independently from the others, without limiting the number of migrants per household and thus including the case of whole-household migration. In my framework, the endogenous sample selection resulting from individual migration may be driven either by whole-household migration or by intra-household selection into migration.

2 The effect of migration on members left behind and the intra-household selection problem

The effect of international migration on the the time allocation (or education) of members remaining in the source country has been explored in various studies. The regression which is generally estimated is as follows :

$$y_{ih} = \alpha + \beta . m_h + u_{ih} \tag{1}$$

[5] For example, among the sample of children living in migrant families, children staying at origin are selected because :(i) their migrating parents have preferred to take other sons/daughters with them (ii) their migrating parents are not rich enough to take any children with them . The first case refers to intra-household selection while the second to whole-migration selection.

where y_{ih} denotes the outcome of individual i living in household h (e.g. labor supply) , m_h is a binary indicating whether at least one family member currently lives abroad and u_{ih} an error term.

The selection problem these studies usually address is the self-selection of households into migration. For example, families facing employment constraints in the domestic labor market may be more likely to send migrants. Lack of employment opportunities and thus low (constrained) labor supply would cause out-migration rather than the reverse [6]. The principal concern is that the error term is correlated with migration, that is $\mathbb{E}[u_{ih}|m_h] \neq 0$. Diverse empirical strategies have been carried out to deal with these endogeneity concerns : selection on observables (Rodriguez and Tiongson, 2001), fixed effect approaches (Mu and van de Walle, 2011) or instrumental variables (Acosta, 2006; Lokshin and Glinskayai, 2008; Binzel and Assaad, 2011; Mendola and Carletto, 2012).

In reality, when estimating equation (1), researchers seek to identify the effect of migration on the outcome (e.g. labor supply) of individual i, *provided that the latter does not migrate himself*. When the individual migrates, his labor supply is not of interest and is not observed in general [7]. Let the dummy d_i denote the migration status of individual i. Each household has two subsequent decisions to make: to engage or not in migration and if so, select which members will be migrating . For individual i three alternatives are possible : either no one moves in the family ($m_h = 0$), or he migrates himself ($m_h = 1, d_i = 1$) and he is left behind by another family member ($m_h = 1, d_i = 0$). The regression which is in fact estimated is

$$y_{ih} = \alpha + \beta . m_h + u_{ih} \qquad \text{observed only if } d_i = 0 \tag{2}$$

The difference in unobservables between individual left behind and those living in non-migrant households can be interestingly decomposed in two terms :

$$\mathbb{E}[u_{ih}|m_h = 1, d_i = 0] - \mathbb{E}[u_{ih}|m_h = 0] = \underbrace{\mathbb{E}[u_{ih}|m_h = 1] - \mathbb{E}[u_{ih}|m_h = 0]}_{\text{inter - household selection}} + \underbrace{P(d_i = 1|m_h = 1) * (\mathbb{E}[u_{ih}|m_h = 1, d_i = 0] - \mathbb{E}[u_{ih}|m_h = 1, d_i = 1])}_{\text{intra-household selection}}$$

with $P(d_i = 1|m_h = 1)$ the conditional probability of migration. The first term corresponds to the well-know self-selection bias *across* households, i.e. the difference in unobservables between households who send a migrant and households who do not. The second term corresponds to the selection bias *within* household, i.e. the difference in unobservables between migrants and left-behinds. This type of bias can be called "intra-household" in the sense that it is driven by the selectivity into which member move and which member stay among households participating in migration.

Implicit in previous approaches is the assumption that $P(d_i = 1|m_h = 1) = 0$, i.e. that individuals under investigation – generally wifes or children– have a probability to migrate equal to zero, in which case there is no "intra-household" selection bias. This hypothesis may be a good approximation in some context. For example, Binzel and Assaad (2011) examines the impact of migration on wifes left in behind

[6] Another example is that the labor supply decisions of female household members may drive the migration decision of the husband. Male migration is likely to depend on whether other household members –including women – are available to help replace the migrants' labor.

[7] Because the survey does not track the migrants

in Egypt, a country where migration is almost exclusively confined to men due to strong social norms. However, in countries like Mexico (or Indonesia) , the individual characteristics of the migrants are very heterogeneous across households. Both males and females migrate, as well as young and older person [8]. Consequently, the distinction between potential migrants and never-migrants– based on observables characteristics – is far from clear-cut. [9]

Various reasons may explain why $\mathbb{E}[u_{ih}|m_h = 1, d_i = 0] \neq \mathbb{E}[u_{ih}|m_h = 1, d_i = 1]$, i.e. why unobservables may differ between migrants and left-behinds. If migration is a long-term collective investment made by the family (Stark and Bloom, 1985) so should decisions as to which members migrate and which stay behind be part of the same welfare-maximizing strategy. In Senegal, Chort and Senne (2013) shows that the migrant is usually the person with the greatest potential of supporting the family in terms of remittances. In Mexico, Antman (2012) and Hernandez-Leon (2008) describe the uneven distribution of responsibility of caring for elderly parents among siblings in the household. Some adult children migrate to the United States an contribute more financially to the parents. Others stay behind and watch over the parents thereby contributing more in terms of time. Migrating sons are the breadwinner of the family. Even if they would have not migrated, they would have likely got a commercial job in the village (or next town) to sustain their parents. In this case , it is likely that $E[u_{ih}|m_h = 1, d_i = 0] < E[u_{ih}|m_h = 1, d_i = 1]$. Naive estimates of β in equation (1) would then be negatively biased, even if households do not self-select in migration. Regarding education outcomes, children or young adult may be positively selected into migration in terms of educational ability , not only across but also within the households. If it is the case, standard estimates of equation (1) would produce downward biased estimates of the causal impact of migration on education.

3 Econometric Setup

I consider the migration of a household member to be the treatment of interest and migration of the individual (whose labor supply is under investigation) to be a "post-treatment complication". The econometric literature usually refers to this sort of problem as endogenous sample-selection (Heckman, 1974). Following the treatment evaluation literature, I use a potential outcome framework (Rubin, 1974). The idea of this approach is to compare the outcome of interest in two hypothetical states of the world: one in which a unit receives the treatment and one in which the same unit does not. For example, we might ask whether a particular individual would participate in the labor market if he lives in a migrant household and whether the same individual would participate if he does not live in a migrant household. Because an individual cannot be in both states of the world (treated and not treated), the fundamental problem of the evaluation is that we cannot observe these two potential outcomes simultaneously.

I define M as the migration status of the *household* , equal to 1 if at least one household member out-migrates and 0 otherwise. I define D as the migration status of the *individual* living in the same

[8] Migration flows to the U.S. are typically composed of young male but women do migrate as well and represent about one third of these flows. See Ibarraran and Lubotsky (2007) – data from Mexican and U.S. censuses

[9] It is of course possible to restrict the analysis on a subsample of individuals who have no chance of migrating either because they are are too young or too old. In the case of Mexico, using the MxFLS survey, this would amount to restrict the sample to men above 50 and women above 40 or to children less than 12. This is certainly not the most interesting population to look at while investigating the effect of migration on labor supply behaviors and time allocation in general.

household , equal to 1 if he out-migrates and 0 otherwise. By definition, $M = 0$ implies $D = 0$. When the family decides to send at least one migrant ($M = 1$), the individual may either be selected as migrant ($D = 1$) or may stay behind at origin ($D = 0$. The decision process might not sequential, but the decision-making unit is the family and not the individual. D depends on M and not vice versa. This family-decision hypothesis has become a generally accepted assumption and has been largely supported by both a theoretical and empirical important body of literature, namely the New Economics of Labor Migration (Lucas, 1997; Stark and Bloom, 1985; Stark and Lucas, 1988; Stark, 1991).

I observe the outcome Y at some point in time after M and D has been realized. In the empirical application , Y is the individual labor supply. Y depends on both the migration status of the household and the migration of the individual. Let $Y(m,d)$ denote the potential values of the outcome. $Y(0,0)$ is the outcome of the individual in case no household member migrates. $Y(1,0)$ is the outcome in case one (or many) family member migrates and the individual stays behind. $Y(1,1)$ is the outcome at destination, in case the individual migrates himself. Similarly, $D(m)$ denotes the potential migration status of the individual as a function of the family's migration decision. $D(1)$ is the individual migration status in case the family to engage in migration. By definition $D(0) = 0$.

In this setting, the difference $Y(1,0) - Y(0,0)$ is the potential effect of migration in case the individual would stay at origin. The only group of people for which the outcome at origin – i.e. $Y(m,0)$ – can be observed under both migration states of the household ($m = 1,0$) is the group for which $D(1) = 0$, i.e. individuals who would never migrate themselves. This is a latent group, distinct from the group of "realized" non-migrants for which $D = 0$. Whether individuals belong to it is unobservable. In particular, $D(1)$ is unobserved for individuals living in non-migrant households $m = 0$. The policy-relevant parameter of interest is therefore the average effect of migration on individuals who would never migrate :

$$\theta = \mathbb{E}[Y(1,0) - Y(0,0)|D(1) = 0]$$

4 Identification with randomly assigned household migration

4.1 Setting

In order to concentrate on the identification problem caused by individual migration D, I assume a random assignment of the migration status of the household. I will relax this assumption in a second step. Random assignment of M means that all potential outcomes are independent of M. However, the actual outcomes Y and D are not independent of M .

Assumption 1 . *Randomly assigned household migration status M*

$$\{Y(1,0), Y(0,0), Y(0,0), D(1)\} \perp M$$

Consider now the the potential migration of the individual . Based on the value of $D(1)$, individuals can be stratified into two latent group (or principal strata). With reference to the Local Average Treatment (LATE) framework (Imbens and Angrist, 1994), I refer to the types defined by $D(1) = 1$ as *potentials*

migrants and to the types $D(1) = 0$ as *Never migrant* (Table 1) . The observed group $\{M = 1, D = 0\}$ is composed of *Never migrant* only, while the observed group $\{M = 0, D = 0\}$ is composed of both *Potential migrants* and *Never migrant* .

Since the observed group $\{M = 1, D = 0\}$ corresponds directly to the latent group N of *Never migrant*, the outcome under treatment for never migrants is directly identified as :

$$\mathbb{E}[Y(1,0)|N] = \mathbb{E}[Y|M = 1, D = 0]$$

The group of non migrating households ($M = 0$) is a mixture of potential migrants and never migrants. The observed outcome is therefore a mixture of the potential outcome of these two latent strata under the no migration regime. Noting p_N the share of never migrant and p_P the share of potential migrants:

$$\mathbb{E}[Y|M = 0, D = 0] = p_N \mathbb{E}[Y(0,0)|N] + p_P \mathbb{E}[Y(0,0)|P]$$

In a non migrant household the econometrician does not know which member is the potential migrant and which member would have stayed home. As a result, $\mathbb{E}[Y(0,0)|N]$ is not point identified. The average effect of migration on never migrants θ is only partially identified. Note that collecting data on migrants to observe the outcome Y of migrants themselves ($D = 1$) would not solve the problem. Note also that this framework encompasses the case of whole-household migration, i.e. migration of all the household members.

A researcher ignoring this selection problem might estimate the difference $\mathbb{E}[Y|M = 1, D = 0] - \mathbb{E}[Y|M = 0, D = 0]$, which equals θ plus a selection term :

$$\mathbb{E}[Y|M = 1, D = 0] - \mathbb{E}[Y|M = 0, D = 0] = \theta + p_P \left(\mathbb{E}[Y(0,0)|N] - \mathbb{E}[Y(0,0)|P]\right)$$

Under assumption 1 , latent group's shares are simply $p_P = P(D = 1|M = 1)$ and $p_N = 1 - p_P$. As underlined in the previous section, the "naive" difference in means correctly estimates θ only if either $P(D = 1|M = 1) = 0$ or $\mathbb{E}[Y(0,0)|N] = \mathbb{E}[Y(0,0)|P]$, in which case there is no intra-household selection of migrants.

Table 1: Principal strata and observed group with randomly assigned household migration

$D(1)$	Latent group	description
1	P	potentials migrant
0	N	never migrant

	D=0	D=1
$M = 1$	**N**	P
$M = 0$	**P,N**	

4.2 Bounds on migration effect

Following Zhang and Rubin (2003) and Lee (2009), sharp bounds for $\mathbb{E}[Y(0,0)|N]$ can be derived. The individual migration status D is equivalent to the sample selection indicator in the framework of Lee (2009) and the household migration status M to the "treatment". Lee's bounds require the assumption that the treatment can only affect sample selection in one direction. In my setting, this monotonicity assumption is satisfied since $D(0) = 0$ by definition : "non treated" individual – i.e. living in non-migrant households – do not migrate and remain in the sample.

The trimming procedure they propose is simple.I describe the procedure for the lower bound as follows. The observed group of non migrating households $M = 0$ is composed of potential migrants C and never-migrants N. In the "worst-case" scenario , the highest potential outcome $Y(0,0)$ of the never-migrants is lower than the lowest outcome of the potential migrants. In this case, we can remove the upper p_P quantiles from the distribution of Y in the group $M = 0$ and estimate the average outcome for the remaining individuals. This gives us the lowest possible outcome for never migrants under control. The upper bound can be derived in similar way, but now trimming the lower tail of the observed outcome distribution.

Let $q(r)$ be the r-quantile of the distribution of $Y|M = 0$.The unknown value $E[Y(0,0)|N]$ can be bounded from above by the mean of Y in the upper $q(1-p_P)$ quantiles from below by the mean in the lower $q(p_P)$ quantiles. Bounds of $E[Y(0,0)|N]$ are:

$$\mathbb{E}^U[Y(0,0)|N] = \mathbb{E}[Y|M = 0, D = 0, Y \geq q(p_P)]$$
$$\mathbb{E}^L[Y(0,0)|N] = \mathbb{E}[Y|M = 0, D = 0, Y \leq q(1-p_P)]$$

and for the causal effect θ:

$$\theta^U = \mathbb{E}[Y|M = 1, D = 0] - \mathbb{E}^L[Y(0,0)|N]$$
$$\theta^L = \mathbb{E}[Y|M = 1, D = 0] - \mathbb{E}^U[Y(0,0)|N]$$

4.3 Can θ be identified using experimental data?

An important question is whether it is possible to identify θ using a randomized control trial (RCT) or experimental data in general. We can imagine the following ideal experimental design. Assume households participates in a visa lottery with only one application per household and with only one visa granted per winning application. Households winning the lottery are however NOT free to choose which family member can use the visa to migrate. Instead, the final individual recipient of the visa is randomly drawn among family members living in the same ballot-winning household. Suppose all individuals receiving the visa ($V = 1$) out-migrate ($D = 1$), i.e. there is full take-up. Suppose also that no ballot looser can migrate. Because of random assignment of visas and perfect compliance, household migration status M is random.

Let $D(m,v)$ denote the individual migration status which is a function of the household migration status M and the visa assignment V. Note that $D(0,v) = 0$ for $v = 0,1$. So I will focus on $D(1,v)$. In this setting , the assignment of visa (V) is a an ideal instrument for D: it is randomly assigned, satisfies the

exclusion restriction and has a a monotone effect on individual migration D .

- $Y(m,d,v) = Y(m,d,v') = Y(m,d) \ \ \forall m,d,v,v' \in \{0,1\}$
- $\{Y(1,0), Y(1,1), Y(0,0), D(1,0), D(1,1)\} \perp V$
- $D(1,1) \geq D(1,0)$

Two latent group of individuals with respect to visa assignment V can be differentiated: always migrants (A) for which $D(1,1) = D(1,0) = 1$ and compliers (C) for which $D(1,1) = 1$ and $D(1,0) = 0$. Always migrants migrate irrespective of whether they win the visa lottery, provided that someone else in the family receives a visa and migrates; compliers migrate only if they receive the visa but do not otherwise. There is no never-migrants because, in this settings, winners of the visa lottery always decide to migrate [10]. As table 2 shows, compliers (C) is the only population for which the the outcome Y at origin can be observed under both migration states of the household ($M = 1$ and $M = 0$). The only interesting parameter is therefore the causal impact of migration among the latent group of compliers : $\theta_C = \mathbb{E}[Y(1,0) - Y(0,0)|C]$.

Table 2: Principal strata and observed group with visa lottery experiment

$D(1,1)$	$D(1,0)$	Latent group
1	1	A , always migrants
1	0	C , complier

		D=0	D=1
$M=1$	$V=1$		A,C
	$V=0$	**C**	A
$M=0$	$V=0$	**A,C**	

The potential outcome $Y(1,0)$ under treatment is directly identified for compliers: $\mathbb{E}[Y(1,0)|C] = \mathbb{E}[Y|M=1,V=0,D=0]$. However, the potential outcome $Y(0,0)$ cannot be identified among compliers (C) because of the presence of always migrants (A) .

The requirement for the identification of θ_C is that individuals loosing the visa lottery ($V = 0$) do not (cannot) migrate, i.e. $D(1,0) = 0$. This assumption can be directly tested in the data since the share of always migrant is simply $p_A = P(D = 1|V = 0, M = 1)$. p_A equals zero only if members loosing the lottery cannot accompany or rejoin the family member who migrates. This is unlikely in a context where legal/physical barriers do not limit clandestine migration and strict rules do not constraint family reunification. Of course θ_C is also identified if $E[Y(0,0)|A] = E[Y(0,0)|C]$, i.e. if always-migrant (A) individuals are not self-selected relative to compliers(C). This assumption is however not testable and unlikely because inconsistent with the literature on self-selection into migration.

4.4 Partial identification as a practical solution

As McKenzie and Yang (2010) reviewed it, experimental approaches in migration studies are scarce. Experiments able to deal with the second form of selectivity into which family members move and

[10] This simplification does not change the result of non-identifiability of the causal migration impact

which remain in the home country are even fewer. To the best of my knowledge, the only exception is the study by Gibson et al. (2011) who exploits a migration lottery in Tonga providing the opportunity for 250 Tongans to move to New Zealand every year. Only one person by household can register for the lottery. This person is the Principal Applicant and if he is successful, his immediate family - spouse and dependent children - can also apply as Secondary Applicant. Importantly, successful applicants cannot take other members of their households to New Zealand(typically parents, siblings or other relatives). As emphasized by the authors, the identification of the causal migration impact crucially relies on the rule specifying which family members can and cannot accompany the successful migrant. The important natural barriers to clandestine migration between Tonga and Auckland (2000 km of sea) ensures that families cannot bypass these rules. Would these policy rules not be effective binding constraint , identification would be questionable [11]

Point identification of the impact of migration on left behind family members is therefore truly demanding. It requires very specific, if not exceptional, policy or field experiments that researchers do not have always at their disposal. Using partial identification instead may be a reasonable solution for migration studies exploring how left behind members are affected by the migration of a relative. In the the remainder of the paper, following Steinmayr (2014) I derive non parametric bounds for the effect of migration on remaining members.

5 Partial Identification with non-random household migration

5.1 Setting

In practice, households self-select into migration and household migration is not random. Empirical studies generally use an instrument for the migration decision of the household – in the Mexico-US migration literature see McKenzie and Rapoport (2007, 2011); Woodruff and Zenteno (2007) . I therefore drop assumption 1 of random assignment of M and assume that a binary instrument $Z \in \{0,1\}$ exists, which is randomly assigned and affects the migration decision of the household. $M(z)$ denotes the potential migration status of the household as a function of the value of the instrument Z. For the moment, let me note also $D(m,z)$ the potential migration of the individual and $Y(m,d,z)$ the outcome as a function of Z. In the presence of a sample selection problem, I have to make additional assumptions compared to the classical IV framework (Imbens and Angrist, 1994). Specifically I make the following assumptions.

Assumption 2 . *Exclusion restriction of Z with respect to Y*

$$Y(m,d,z) = Y(m,d,z') = Y(m,d) \;\; \forall m,d,z,z' \in \{0,1\}$$

Assumption 2 states that the effect of Z on the potential outcomes Y must be via an effect of Z on M and D. To put it differently, the instrument may impact the labor supply of family members only only

[11] To cite the authors:" We use the age and relationship rules governing which Secondary Applicants can move with the Principal Applicant to identify household members that would have moved to New Zealand if the Principal Applicant had been successful and compliant with the treatment. These rules appear to be the binding constraint since the remaining family of PAC emigrants are almost all outside the age and relationship eligibility for moving to New Zealand"

through its effect on the migration of the household members. In addition, I assume that the instrument is randomly assigned and therefore independent of all potential outcomes:

Assumption 3 . *Randomly assigned instrument Z*

$$\{Y(m,d), M(z), D(m,z)\} \perp Z \ \ \forall m,d,z \in \{0,1\}$$

I now distinguish different latent strata with respect to the *instrument*. I can differentiate among *households* between the Always takers(A), the Compliers (C), the Defiers (D) and the Never takers (N). Always takers are households who would send a migrant irrespective of the value of theinstrument ; compliers send a migrant if the instrument takes on the value of one but not if it takes on the value of zero; defiers migrate if the instrument equals zero but not if the instrument equals one ; and never taker never migrate irrespective of the value of the instrument.

Individual migration depends on both M and Z, but since M is itself a function of the instrument, D is simply a function of Z, i.e. $D = D(M(Z), Z) = D(Z)$. I can distinguish between four different types of *individuals* defined with respect to the *instrument* : always migrant (A) for which $D(1) = D(0) = 1$; compliers (C) defined by $D(1) = 1$ and $D(0) = 0$; defiers (D) for which $D(1) = 0$ and $D(0) = 1$; and never migrants (N) defined by $D(1) = D(0) = 0$. Combining the four strata of individuals with the four strata of households gives in total $4 \times 4 = 16$ latent strata (see table 9 in Appendix A1). We refer to the strata using a two letter system, the first letter indicating the type of household , the second the type of individual. E.g. CN refers to never migrating individuals in compliers household. Since $M = 0$ implies that $D = 0$ by definition, this is also true for potential migration decisions at a given value of Z [12]. The fact thatfor all $z \in \{0,1\}$ $M(z) = 0 \Rightarrow D(z) = 0$ rules out the existence of strata NA, NC, ND , CA, CD, DA,DC [13].

I now make two additional assumption that are common in the LATE literature. First, I assume a monotone effect of the instrument on the household migration M. This assumption states that every household is at least as likely to send a migrant if $Z = 1$ as it should be with $Z = 0$. It thus rules out the existence of defiers (D) among households.

Assumption 4. *Monotonicity of M in Z (no defiers)*

$$M(1) \geq M(0)$$

Second, I also assume a non-zero average effect of Z on the migration decision of the household. This assumption amounts to ensure the existence of the latent group of compliers (C) among households:

Assumption 5. *Non-zero effect of Z on M (existence of compliers)*

$$\mathbb{E}[M(1) - M(0)] > 0$$

I now make an important assumption specific to this framework: I suppose that the effect of the instrument on the potential migration of the *individual* must be via an effect on M. To put differently,

[12] there is no potential migrant individuals in never migrating household

[13] In strata CA and CD, $m(0) = 0$ implies that $D(0) = 0$ and in strata NA, NC,ND $m(1) = m(0) = 0$ implies $d(1) = d(0) = 0$

the decision of the family as to which member(s) migrate should not be influenced by the value of the instrument. The instrument is supposed to have no effect on the intra-household selection of the migrant individuals. In a later step I will derive alternative bounds in case this assumption is violated and replace this assumption by another monotonicity hypothesis.

Assumption 6. *Exclusion restriction of Z with respect to D*

$$D(m,z) = D(m,z') = D(m) \;\; \forall m,z,z' \in \{0,1\}$$

Assumptions 6 rules out the existence of strata AC and AD. In these strata , the household migration status does not react to the instrument (because $M(1) = M(0) = 1$) while the individual migration is affected by Z (since $D(1) \neq D(0)$). The instrument has thus a direct effect on D in these strata, which is precisely what Assumption 6 excludes. Table 3 shows the correspondence between observed group and latent strata that remain after Assumptions 4 and 6 are imposed.

The only latent group for which the outcome at origin $Y(m,0)$ can be observed under both migration states of the household is the stratum CN. In this latent group, the instrument induces the family to send at least one migrant but the individual does not migrate himself . The causal migration impact for this stratum is therefore the local average effect of migration for individuals who never migrate [14]. In the rest of the section, I will focus on the partial identification of the causal effect :

$$\theta_{CN} = \mathbb{E}[Y(1,0) - Y(0,0)|CN]$$

Table 3: Latent and observed groups with non random household migration M and exclusion restriction $D(M,Z) = D(M)$

$M(1)$	$M(0)$	$D(1)$	$D(0)$	Latent group
1	1	1	1	AA
1	1	0	0	AN
1	0	1	0	CC
1	0	0	0	CN
0	0	0	0	NN

		D=0	D=1
$Z=1$	$M=1$	**AN, CN**	AA, CC
	$M=0$	**NN**	
$Z=0$	$M=1$	**AN**	AA
	$M=0$	**CC, CN ,NN**	

5.2 Bounds on the migration effect

Under the assumptions 2,3,4,5 and 6 , only five different latent groups remain. Six different groups are observed based on realized D and M across Z. In total, four unknown proportions can be identified

[14] It is a local effect in the sense that it is identified only for the population of households whose migration decision is affected by the instrument

from four linearly independent equations. [15] To simplify the notation , I denote $\overline{Y}^{zmd} = \mathbb{E}[Y|Z=z,M=m,D=d]$ for the observed average outcome in the observed group $\{Z=z,M=m,D=d\}$. The potential outcome under treatment $Y(1,0)$ for the latent group CN is observed as part of the mixture distribution in the observed group $\{Z=1,M=1,D=0\}$. In addition, the outcome under treatment for the latent group AN is directly observed in the observed group $\{Z=0,M=1,D=0\}$. Combining these two remarks allows identifying the expected outcome under treatment for CN :

$$\mathbb{E}[Y(1,0)|CN] = \frac{(p_{AN}+p_{CN})\overline{Y}^{110} - p_{AN}\overline{Y}^{010}}{p_{CN}} \tag{3}$$

I follow Chen and Flores (2012) and Steinmayr (2014) to derive bounds for the potential outcome of CN under control $\mathbb{E}[Y(0,0)|CN]$. In appendix A2 , I detail and explain the derivation of the non-parametric lower and upper bounds $\mathbb{E}^L[Y(0,0)|CN]$ and $\mathbb{E}^U[Y(0,0)|CN]$. Bounds for the causal effect can be constructed by combining these bounds with the point identified potential outcome under treatment for the latent group CN:

$$\begin{aligned} \theta^U &= \mathbb{E}[Y(1,0)|CN] - \mathbb{E}^L[Y(0,0)|CN] \\ \theta^L &= \mathbb{E}[Y(1,0)|CN] - \mathbb{E}^U[Y(0,0)|CN] \end{aligned} \tag{4}$$

5.3 Bias in the standard LATE estimator

5.3.1 Asymptotic bias

It is interesting to compare the standard LATE (Local Average Treatment Effect) with the parameter of interest θ_{CN}. A researcher neglecting the intra-household selection would estimate the LATE on the sample of individuals who stay at origin and for whom the outcome Y is observed.

$$LATE \equiv \frac{\mathbb{E}[Y|Z=1,D=0] - \mathbb{E}[Y|Z=0,D=0]}{\mathbb{E}[M|Z=1,D=0] - \mathbb{E}[M|Z=0,D=0]}$$

After some computations , it can be shown that the standard LATE converges towards θ_{CN} plus three selection terms:

$$\begin{aligned} LATE = \theta_{CN} + \frac{p_{CC}}{\pi}\Big[&(p_{AN}+p_{CN})\left(\mathbb{E}[Y(0,0)|CN] - \mathbb{E}[Y(0,0)|CC]\right) + p_{NN}\left(\mathbb{E}[Y(0,0)|NN] - \mathbb{E}[Y(0,0)|CC]\right) \\ &+ p_{AN}\left(\mathbb{E}[Y(1,0)|AN] - \mathbb{E}[Y(1,0)|CN]\right)\Big] \end{aligned} \tag{5}$$

with $\pi = p_{CN}(1-p_{AA}) + p_{AN}p_{CC}$. Two important remarks must be made. First, if in households who

[15]

$$\begin{aligned} p_{NN} &= P(M=0,D=0|Z=1) \\ p_{AN} &= P(M=1,D=0|Z=0) \\ p_{CN} &= P(M=1,D=0|Z=1) - P(M=1,D=0|Z=0) \\ p_{CC} &= P(M=0,D=0|Z=0) - p_{CN} - p_{NN} \\ p_{AA} &= P(M=1,D=1|Z=0) = 1 - p_{CC} - p_{CN} - p_{AN} - p_{NN} \end{aligned}$$

comply with in instrument individuals have a zero probability to migrate then $p_{CC} = 0$ and the standard Wald estimate consistently converge towards θ_{CN} [16]. This might the case only if the econometrician is able to restrict the sample to individuals, who, based on their observables characteristics, are unlikely to migrate themselves (e.g. women in Egypt) but may be left behind by other potential migrant members who are excluded from the sample (e.g. men in Egypt).

Second and most importantly , the LATE may be inconsistent even in the absence of the second form of selectivity due to the intra-household selection of the migrants. Even if migrants do not self-select within complier households, i.e. if $\mathbb{E}[Y(0,0)|CN] = \mathbb{E}[Y(0,0)|CC]$, LATE may be asymptotically biased. The LATE estimator is prone to bias because it does not take into account the fact that the instrument generates an endogenous sample selection. Among the group for which $Z = 1$, complier individuals living in complier households (CC) migrate and their outcome is no more observed at origin. In contrary, among the group for which $Z = 0$, all individuals have stayed at origin, despite (CC) types would have migrated in case they would have received the instrument $Z = 1$. This discrepancy in the proportion of the latent group (CC) between the group $Z = 1$ and $Z = 0$ suffices to generate a bias – even when the potential outcome are identical between the (CC) and (CN) group.

Of course, a sufficient condition for which LATE is unbiased is that $\mathbb{E}[Y(0,0)|CN] = \mathbb{E}[Y(0,0)|CC] = \mathbb{E}[Y(0,0)|NN]$ and $\mathbb{E}[Y(1,0)|AN] = \mathbb{E}[Y(1,0)|CN]$. However, in this case, there is no selection at all, not even ac cross households. The use of an instrument is pointless and a simple difference in means $\mathbb{E}[Y|M = 1] - \mathbb{E}[Y|M = 0]$ consistently estimates θ_{CN}

Finally, as I show in appendix A4 , the LATE estimator can possibly lie outside the non-parametric bounds a la Chen and Flores (2012) derived in the previous section in equation (4) . Examples can be build in which the LATE is above the upper bound θ^U.

5.3.2 Adjusted IV estimator in the absence of intra-household selection

More generally, if there is no systematic intra-household selection into migration among compliers households , i.e. if $\mathbb{E}[Y(0,0)|CN] = \mathbb{E}[Y(0,0)|CC]$, then the causal migration impact θ_{CN} among the (CN) population can be identified by :

$$\widehat{\theta_{CN}}^{IV} = \frac{(p_{AN} + p_{CN})\overline{Y}^{110} - p_{AN}\overline{Y}^{010}}{p_{CN}} - \frac{(p_{NN} + p_{CN} + p_{CC})\overline{Y}^{000} - p_{NN}\overline{Y}^{100}}{p_{CN} + p_{CC}} \tag{6}$$

Note that $\widehat{\theta_{CN}}^{IV}$ is generally different from the LATE estimator. The bias of the LATE can be expressed as :

$$LATE - \widehat{\theta_{CN}}^{IV} = \frac{p_{CC}}{\pi}\left[p_{NN}(1 + \frac{p_{NN}}{p_{CC} + p_{CN}})(\overline{Y}^{100} - \overline{Y}^{000}) + \frac{p_{AN}(p_{AN} + p_{CN})}{p_{CN}}(\overline{Y}^{010} - \overline{Y}^{110})\right]$$

[16] $p_{CC} = P(D = 1|Z = 1) - P(D = 1|Z = 0)$

5.4 Alternative bounds without assumption 6

Assumption 6 may be debatable in some context. For instance, migration networks at destination may certainly be used as instrument because they reduce the cost of migration and thereby affect the family decision as to whether send migrant(s) or not. However, such networks may also influence family decision as to which and how many members are sent. If it is the case, assumption 6 would be violated.

I thus drop assumption 6 in this subsection. Now the existence of the latent strata AC and AD cannot be ruled out anymore. In these latent groups, the instrument has a direct effect on individual migration. I then make the alternative assumption that the effect of the instrument on *individual* migration is monotonic: every individual is more likely to migrate if $Z = 1$ than if $Z = 0$. This excludes the existence of defiers and therefore the existence of the stratum AD [17]

Assumption 7. *Monotonicity of D in Z*

$$D(1,1) \geq D(1,0)$$

The proportion of the latent groups are no more identified.Indeed, there are five unknown strata proportions for only four known probabilities (linear independent equations) [18]. I therefore introduce an additional parameter λ, which equals the ratio of the share of always migrant individuals to the the share of complier individuals in always migrant households :

$$p_{AC} = \lambda p_{AA}$$

The sensitivity of the bounds with respect to the value of $\lambda \geq 0$ will be explored in the empirical application. We know that $p_{AA} = P(D = 1|Z = 0)$ that $P(M = 1, D = 0|Z = 0) = p_{AN} + \lambda p_{AA}$ and also that $P(D = 1|Z = 1) = p_{CC} + (1 + \lambda)p_{AA}$. These three equations and the requirement that $p_{AN} \geq 0$ and $p_{CC} \geq 0$ give a interval of possible values of λ, i.e.

$$0 \leq \lambda \leq \lambda_{max} = \min(\lambda^{an}_{max}, \lambda^{cc}_{max})$$

$$\text{with } \lambda^{an}_{max} = \frac{P(M = 1, D = 0|Z = 0)}{P(D = 1|Z = 0)} \text{ and } \lambda^{cc}_{max} = \frac{P(D = 1|Z = 1)}{P(D = 1|Z = 0)} - 1$$

I use the procedure of Huber and Mellace (2013) to derive sharp bounds on θ_{CN} which depends on the value of λ. The formulas and details of the derivation are provided in the appendix A3.

[17] As the rest of the paper will show, this assumption only affects the bounds for $\mathbb{E}[Y(1,0)|CN]$: they are less tight if the existence of defiers is not rule out

[18] The 6th strata proportions is one minus the sum of the others. There are only 4 independent equations because

$$P(M = 1, D = 1|Z) + P(M = 1, D = 0|Z) + P(M = 0, D = 0|Z) = 1$$

6 Estimation and inference

6.1 Estimating migration probabilities : the issue of whole-household migration

As underlined by Steinmayr (2014), the estimation of the probability to migrate are problematic when using cross-sectional data. By construction, cross-sectional surveys do not include households where all members have migrated as no household member is left to respond to the survey. When estimating migration probabilities, one should normally takes into account this type of whole-household migration. Using a Mexican cross-sectional survey, Steinmayr (2014) proposes a method to correct the migration probabilities (to the U.S.) by exploiting discrepancies in the number of Mexican between the U.S. and Mexican census.

Because I use a panel survey, I do not face similar difficulties. In the empirical application, I use the two rounds of the Mexican Family Life Survey conducted in 2002 and 2005. As described in Rubalcava and Teruel (2008) and Teruel et al. (2012) when a person moved and was not found in the same household of origin (at the baseline survey) , enumerators inquired about his/her whereabouts by asking members left behind in the original household about his/her new location. In cases where the whole household moved, respondent's friends, relatives or neighbors provided the location of the absent household. Hence, even if they could not be individually recontacted, all migrants to the U.S. can be identified, irrespective of whether they leave behind household members or not.

In the rural sample of the MxFLS life survey , I find that 4% of working-age respondents (15-65) have migrated to the U.S. between 2002 and 2005. The attrition rate at the household level – i.e. percentage of households that are not re-interviewed in 2005 in Mexico – is low, of about 3% . Among these "attriter" households, I find that 20% have actually migrated as a whole to the U.S.

Based on the 2000 U.S. census, 1,492,111 Mexican immigrants live in the U.S. Based on the 2000 Mexican census, only 1,221,598 Mexicans have migrated to the U.S. [19]. The discrepancy corresponds to all-move households who migrate as a whole and are not counted in the Mexican census. Based on theses figures , migrants moving with their entire family represent $\frac{1{,}492{,}111-1{,}221{,}598}{1{,}492{,}111} = 18\%$ of all U.S. migrants, a proportion close to what I find using the MxFLS, namely $\frac{48}{328} = 15\%$ .This suggests the MxFLS does not miss many cases of whole-household migration.

Table 4: Migrants in the rural sample of working wage individuals - MxFLS 1-2 (absolute frequencies)

	"Attriter" household		
	No	Yes	Total
Non migrant	6,963	184	7,147
U.S. migrant	280	48	328
Total	7,243	232	7,475

[19] See Ibarraran and Lubotsky (2007) estimates. The size of the Mexican immigrant population living in the U.S. is computed using two different data sources: (i) the 2000 U.S. census, which is supposed to provide the exhaustive number of migrants (who did not return) (ii) the 2000 Mexican census, which gives the number of migrants who leave behind at least one household member in Mexico. Left-behind report whether any household member has migrated in the U.S. in the last five years. The U.S. census sample includes only people who report they came to the United States between 1995 and 2000 .

6.2 Using covariates to narrow bounds and weaken assumption 3

The use of exogenous covariates - such as baseline socio-demographics characteristics - can serve two different purposes. First, as Lee (2009) shows, the use of covariates tightens the bounds. Bounds are narrower when using baseline characteristics than when not using it. Second, the assumption of random assignment of the instrument Z might be valid only conditionally on a set of covariates X. We might then want to replace assumption 3 of unconditional independence by a weaker conditional independence assumption.

To identify bounds conditional on X, I need to make new assumptions and to modify some of the previous hypothesis. My framework is quite different as Lee (2009) because I use an instrument Z and because I do not assume that the instrument is orthogonal to the X. My econometric setup is closer to Frolich (2007) who proposes a non parametric estimation of the LATE with covariates. I adapt his procedure to my problem. I begin by assuming that my covariates are exogenous :

Assumption 9. *Exogenous covariates*

$$X_i(m,d,z) = X_i \ \ \forall m,d,z \in 0,1$$

where $X(m,d,z)$ is the potential value of X for unit i that would be observed if M, D and Z were set by external intervention. This assumption precludes that X itself is caused by the migration or by the instrument. However it does not forbid X to affect the probabilities of migration or to determine the instrument. Now I weaken assumption 3 of random assignment of Z by allowing the instrument to be unconditionally correlated with the potential outcomes (and potential migration) . I make a standard conditional independence assumption :

Assumption 3* . *Randomly assigned instrument Z conditionally on X*

$$\{Y(m,d), M(z), D(m,z)\} \perp Z \,|X \ \ \forall m,d,z \in \{0,1\}$$

Finally, I suppose that the support of X is the same for the to subpopulation of $Z=1$ and $Z=0$. This hypothesis ensures that a local average treatment effect conditional on $X=x$, as well as its bounds, are well defined for all x :

Assumption 10 . *Common support*

$$0 < P(Z=1|X=x) < 1 \text{ for all } x \text{ with positive density}$$

Under assumption 2 (exclusion restriction with respect to Y) , 3* , 4 (no household defiers) , 5 (existence of compliers) , 6 (exclusion restriction with respect to D) , 9 and 10, an upper and lower bound of θ_{CN} can be constructed in each cell $X=x$. The same procedure as before can be applied conditional on X, i.e. stratified by observed characteristics. Then by averaging across the distribution of X conditional on CN, we can obtain sharp lower and upper bounds for θ_{CN}. Assume that each element of the vector of covariates X has a discrete support so that this vector can take on one finite number of discrete values $\{x_1, x_2, ..., x_J\}$. Let $p(x_k)$ denote the proportions of individuals with characteristics x_k. Then note also

with reference to section 3.2 and equation (4)

$$\theta^U(x) = \mathbb{E}[Y(1,0)|CN, X = x] - \mathbb{E}^L[Y(0,0)|N, X = x]$$
$$\theta^L(x) = \mathbb{E}[Y(1,0)|CN, X = x] - \mathbb{E}^U[Y(0,0)|N, X = x]$$

Proposition 1 (adapted from Lee (2009)) *Under assumptions 2,3*, 4,5,6,9 and 10 ,* $\overline{\theta^U}$ *and* $\overline{\theta^L}$ *are sharp lower and upper bound for the average treatment* $\theta_{CN} = \mathbb{E}[Y(1,0) - Y(0,0)|CN]$ *where:*

$$\overline{\theta^U} = \sum_{j=1}^{J} \theta^U(x_j) * P(X = x_j|CN)$$
$$\overline{\theta^L} = \sum_{j=1}^{J} \theta^L(x_j) * P(X = x_j|CN)$$
$$P(X = x|CN) = \frac{p_{CN}(x) * p(x)}{\sum_{k=1}^{J} p_{CN}(x_k) * p(x_k)}$$
$$p_{CN}(x) = P(M = 1, D = 0|Z = 1, X = x) - P(M = 1, D = 0|Z = 0, X = x)$$

In a given cell $X = x$, the bounds of $\theta_{CN}(x) = \mathbb{E}[Y(1,0) - Y(0,0)|CN, X = x]$ are $\theta^U(x)$ and $\theta^L(x)$. If $p_{CN}(x) = 0$ these bounds are not identified. But for the identification of the bounds on all individuals in the latent group CN(compliers never migrants), the assumption 5 that $p_{CN} > 0$ suffices because any value x with $p_{CN}(x) = 0$ receives zero weight in the weighted average . The bounds $\overline{\theta^U}$ and $\overline{\theta^L}$ are sharp in the sense that they are respectively the smallest upper bound and largest lower bound that are consistent with the data. Furthermore, $\overline{\theta^U} \leq \theta^U$ and $\overline{\theta^L} \geq \theta^L$ because more information is used when using covariates.

When assumption 6 is replaced by assumptions 7 and 8, the procedure is similar except that the conditional bounds in each cell $X = x$ are different.

6.3 Discrete outcomes: issue with quantiles

When the outcome Y is discrete – such as participation in the labor market in the empirical example – the occurrence of mass points with equal outcome values entails a non unique quantile function. As suggested in Kitagawa (2009) and Huber and Mellace (2013) , I replace the non-unique quantile function with a rank function in order to break ties. For example, the estimate of the upper tail trimming function $\mathbb{E}[Y|M = m, D = d, Z = z, Y \leq q(p)]$, where $q(p)$ denotes the $p-th$ quantile, can be obtained as follows. I simply sort the observations by increasing order of Y in the observed cell $\{M = m, D = d, Z = z,\}$, giving an (arbitrary) different rank for the observations with the same outcome value. I then estimate the mean in the subsample of the first $p * n$ observations, where n denotes the number of observations in this cell. For deriving the lower tail trimming function $\mathbb{E}[Y|M = m, D = d, Z = z, Y \geq q(1-p)]$, I estimate the mean in the subsample of the last $p * n$ observations.

7 Empirical application: migration and labor supply in Mexico

A growing empirical literature has studied the effect of migration on the labor supply of family members left behind in source country , and especially in Mexico (Amuedo-Dorantes and Pozo, 2006; Hanson, 2007). In Murard (2013), using the rural sample of the Mexican Family Life Survey I examine how the non-migrant individuals' participation (and hours) in different activities – such as non non-agricultural wage work or self-employment – are affected by the migration of a household member to the United States. Using instrumented difference-in-differences estimators (IV with individual fixed effect) , I find that left-behinds reduce their participation in non-agricultural wage work but also increase their self-employed work in response to the international migration of a family member in the U.S. . Importantly, I find that this re-allocation of labor is particularly significant among young women (below 36), who are typically the daughters of the household head.

An important question is whether these results can be interpreted as causal or whether they are biased by intra-household selection of the migrant(s). The division of family role within the household could indeed totally account for these results, even in the absence of any causal effect of migration. For example, in a family with several daughters, parents likely assign different role to their daughters. The (unique) migrant daughter could be the one in which the family has invested the more in terms of human capital ; even if she would not have migrated, she would have stopped farm work and got a commercial job while the other never-migrating daughters are in charge of working in the farm and taking care of the elderly and children [20].

In this paper I will use the same Mexican Family Life Survey as my previous work and estimate non parametric bounds for the effect of migration on remaining young women in Mexican rural household. The narrowness of the bounds will suggest the extent to which previous estimates in Murard (2013) could have been biased. It will also give a confidence interval of the "causal" effect of migration (assuming the validity of the instrument).

7.1 Data

The Mexican Family Life Survey (MxFLS) is a longitudinal household survey representative at the rural level. The baseline survey was conducted from April to July 2002 and collected information from a sample of approximately 3,300 households (14,000 individuals) residing in 75 rural communities with less than 2,500 inhabitants (defined as rural areas) . The second round of the survey was begun in mid-2005 and completed in 2006.

I use a sample consisting of 1521 women between 15 and 36. I define household migration, the treatment, as the fact to live in a household where at least one member has migrated to the U.S. between the two survey rounds, i.e. 2002 and 2005. I define individual migration as the fact to be one of the migrant(s) of the family. In the sample, 86% of women live in a household which does not send any migrant(s) Among the 14% who live a a migrant household, 5% migrate themselves to the U.S. and 9%

[20] De Janvry and Sadoulet (2001) have shown the importance of off-farm activities in the Mexican ejido sector, i.e. peasant communities containing the majority of the rural population and half the country's agricultural land. Non agricultural employment is very frequent and highly varied ; it is typically composed of construction , manufactures, commerce jobs . I find in the MxFLS survey that about 20% of young women work in non agricultural jobs, a figure close to De Janvry and Sadoulet (2001)

are left behind by another migrating family member.

The two outcomes of interest are the participation in the non agricultural labor market and the participation in self-employed activities – mainly farming or micro-businesses in rural villages of Mexico. More precisely, I use the longitudinal structure of the data to apply first time-differences in order to wipe out time-invariant unobservables factors. My outcomes are therefore the *variation* of the labor supply before and after migration, that is between 2002 and 2005 . These variations over time equal to -1 in case the woman stops working , 0 if she keeps working (or not working) and 1 if she starts working. At baseline, I find that about 26% of women work, either in non agricultural jobs (15%) or in self-employment (8%). From 2002 to 2005, I find that more than 25% of women have either switched of activity or stopped working or entered the labor force.

7.2 Evidence of intra-household selection before migration

The longitudinal structure of the data allows to observe the situation of the households before migration has occurred. Before turning to the impact of migration, an interesting question is thus whether, in 2002 at baseline, would-be migrants who will have migrated by 2005 have already a systematically different labor supply behavior than non-migrants. More precisely, among young women living in migrant-sending households, some migrate themselves and some stay behind. Significant differences in the *initial levels* of labor force participation between these two populations would provide suggestive evidence of intra-household selection. Formally, I estimate the following regression on the sample of young women living in households who have sent at least one migrant in the U.S. by 2005 :

$$L_{ih,2002} = \alpha + \gamma.D_{i,02-05} + \beta.X_{ih,2002} + \varepsilon_{ih}$$

where $L_{ih,2002}$ is the initial participation in the labor force (binary), $D_{i,02-05}$ is a binary indicating whether women i has migrated by 2005, and $X_{ih,2002}$ a vector of households and individual characteristics at baseline.[21] Table 10 in Appendix A1 shows that would-be migrant women are initially about 15% more likely to work (either in wage-earning occupation or in self-employment) than women staying behind. This difference remains significant (at 10%) when household fixed effects are included in the regression (column 3). This difference seems to be especially driven by a higher participation in non-agricultural jobs. This finding suggests that migrant women are indeed likely to be in charge of sustaining the family financially while women staying behind might be responsible for tacking care of the elderly and children as well as doing various household chores.

7.3 Past municipal migration networks to instrument for household migration

Unobserved shocks between 2002 and 2005 may affect both migration decisions and labor outcomes. For example, local labor demand shocks may provoke involuntary unemployment in wage-earning jobs and force some members of the family to out-migrate to find jobs elsewhere. The migration of one household member may also reflect joint decisions with family labor allocation : women's participation

[21] age, individual education, household size, number of elderly, number of children under 12 , highest educational level attained in the family, initial social and private transfers received, and initial wealth of the household

in agricultural work may help finance men's out-migration.

To overcome the problem of households self-selection into migration , a number of studies (McKenzie and Rapoport, 2011; Woodruff and Zenteno, 2007) have used historical state-level migration rates as an instrument for current migration levels. Following these studies, I use as instrument the emigration rate to the U.S. from 1995 to 2000 in each Mexican municipality. To derive the migration rates I use the 2000 Mexican Census which records the international migration of each household member during the 5 years prior the interview. This instrument is meant to proxy for the extent of village level migration networks which likely reduce migration costs – such as travel costs (smugglers), initial setup and job search costs at destination.

The exclusion restriction is that these 1995-2000 municipal migration rates do not affect the variation in the labor supply outcomes between 2002 and 2005, expect through current migration of household members. A detailed discussion of this instrument and the exclusion restriction can be found in Murard (2013). Assumption 6 also requires that the instrument does not influence the migration decision of the individual directly, but only the decision of the household as to whether or not send a migrant in the U.S.. This might be a reasonable assumption if, in a context of binding liquidity/credit constraints, migration networks primarily helps to reduce upfront migration costs without affecting the within-family-selection of the migrant individuals.

Finally, I recode the continuous measure of municipal migration rate into a binary variable. I define municipalities as low-migration municipalities ($Z = 0$) if the migration rate is below the sample median (close to 2%) and as high-migration municipalities ($Z = 1$) if migration rate is above the 75th percentile (around 5%). Observations with emigration rate between the two endpoints are neglected. I do this to allow stratification on instrument assignment and to estimate a LATE on the largest population of compliers – see Frolich (2007) for a justification [22].Figure 3 in Appendix A1 shows the relation between emigration rates and the probability of living in a U.S. migrant household at baseline.

7.4 Results

I bound the effect of living in a migrant household on participation in non-agricultural jobs and self-employment among young women below 36 in Mexico. More precisely, the population for which the effect is of interest – and for which bounds can be identified – is the latent group CN of young women. This group corresponds to those young female who would never migrate themselves but who live in a household where the migration of another family member is induced by the availability of community networks in the U.S.. Ignoring the endogenous intra-household selection of the migrant, I estimate the local average effect (LATE) using a simple linear IV estimator without covariates – i.e. a Wald estimator.

[22] Consider a multivalued discrete instrument $Z \in z_0, z_1, z_2$ with $z_2 > z_1 > z_0$. The monotony assumptions requires that $M(z_2) \geq M(z_1) \geq M(z_0)$ Note Compliers $C_{01} = \{M(z_1) = 1, M(z_0) = 0\}$ and Compliers $C_{12} = \{M(z_2) = 1, M(z_1) = 0\}$ The Average treatment on both compliers group is :

$$\mathbb{E}[Y^1 - Y^0 | C_{12} \bigcup C_{01}] = \mathbb{E}[Y^1 - Y^0 | M(z_2) = 1, M(z_0) = 0]$$

Generally if $Supp(Z) = (z_{min}, z_{max})$, the parameter of interest is

$$\mathbb{E}[Y(1,0) - Y(0,0) | M(z_{max}) = 1, M(z_{min}) = 0, D(z) = 0 \; \forall z]$$

The estimated impact is a (statistically) significant increase in self-employment by 27 percentage points and a significant reduction in non-agricultural wage labor by 23 percentage points (table 5).

7.4.1 Bounds without covariates

I begin by estimating in table 5 the bounds derived in equation (4) under assumptions 2,3,4,5, 6 . The expected participation in self-employment under treatment, i.e. in case the household sends a U.S. migrant , is 0.026 for the group CN . The lower and upper bounds of the outcome under control, i.e. in case of no migration, are -0.37 and 0.10 for the group CN. In consequence, the lower and upper bounds for the average effect of migration for the CN stratum are -0.078 and 0.399). The estimated bounds suggest that migration may have a negative effect on self-employment instead of a positive one. Relative to never-migrant daughters, migrant daughters might have been much more likely to stop farming (or less likely to start farming) if they have would not migrated. This type of selection would cause a upward bias of the standard Wald estimator.

With respect to non agricultural wage labor, lower and upper bounds for the average effect are -0.60 and 0.22, suggesting that that migration may have a positive effect. Again, the fact that migrant daughters would have been much more likely to find a local non-rural job than never-migrant daughters can totally account for a downward bias of the standard IV estimate.

In table 5, I also report $\widehat{\theta_{CN}}^{IV}$, the consistent estimator of θ_{CN}derived in equation (6) under the assumption of absence of intra-household selection into migration . It is apparent that the standard LATE overestimates the impact of migration on self-employment, even if there is no selection into migration within complier households. However, the bias in the LATE, measured by the difference between the LATE and $\widehat{\theta_{CN}}^{IV}$,is not statistically significant.

7.4.2 Bounds with covariates

Using panel data, I observe characteristics of individuals and households in 2002 before migration (age,sex,education...) . I now use the baseline characteristics to tighten the bounds for the causal migration impact. However, because the estimation method in proposition 1 is non-parametric, I must first derive the bounds in each population cell defined by a vector of discrete value of characteristics $(x_1, x_2, ..., x_K)$. Because the sample size is not very large and the number of characteristics K is high, I face the well-known problem of "curse of dimensionality" [23]. To bypass this problem, I project the vector of characteristics X on a single one-dimensional index. As a relevant index, I choose the ex-ante propensity to migrate which I estimate using a standard probit model:

$$P(D = 1|X) = P(X\beta + \varepsilon > 0) = \Phi(X\beta)$$

I use $\Phi(X\widehat{\beta})$ to create three population groups of equal size : individuals with low $\widehat{p}(X)$ in the first tercile of the sample distribution , medium $\widehat{p}(X)$ in the second tercile, and high $\widehat{p}(X)$ in the last tercile. I then estimate bounds in proposition 1 using this categorical variable which divides the sample only

[23]The number of observations within each cell shrink very rapidly and the common support condition (assumption 10) is violated

exploiting variation in the exogenous X. Table 11 in appendix A1 shows the estimation of $\Phi(X\widehat{\beta})$ using the initial demographic composition of the households, the initial wealth and nonlabor income, the age , education and marital status of the individual, as well as some characteristics of the municipality.

When I use these covariates in table 6, , I obtain tighter bounds as expected – also see table 12 in appendix A1 for details. With respect to self-employment, I find lower and upper bounds on θ_{CN} of 0.247 and 0.377 , close to the LATE estimate. This interval lies strictly within the one in the previous table 5 without covariates. Since the lower bounds is positive, it appears that the causal effect of migration is positive, even when the endogenous intra-household selection of migrants is taken into account. Migration seems to cause young women left behind to increase their participation in self-employed activities. With respect to non-rural jobs, I obtain a lower bound of -8% suggesting that the true magnitude of the causal impact of migration θ_{CN} may be lower than what the LATE indicates. However, the lower bound is negative which indicates that the causal migration impact θ_{CN} is strictly negative. Migration seems to cause a decline in the participation in non-rural jobs among young women staying behind. Overall, non parametric bounds on θ_{CN} tend to confirm the findings of a re-allocation of labor away from non-agricultural jobs to farm work and other self-employed activities.

7.4.3 Bounds without assumption 6 : sensitivity with respect to λ

By reducing the cost of migration, migration networks in the U.S. probably influences the decision of the family as to whether or not send migrant(s) . However, networks might also affect the decision as to which and how many members are sent. For example, by facilitating initial setup at destination, the availability of kinship/community networks in the U.S. might determine whether girls/daughters accompany the principal male migrant. If it is the case, assumption 6 of exclusion restriction of Z with respect to D, i.e. $D(m,z) = D(m)$ might be violated.

As an alternative,I derive bounds under Assumption 7 instead of Assumption 6. Assumption 7 allows for a direct effect of migration networks on individual migration but requires this effect is to be monotonic. Since networks lower the cost of migration, they likely have a monotonic positive effect on individual migration. Assumption 7 also introduces an additional parameter λ, the ratio of the proportion of compliers individuals to the proportion of alway-migrant individuals in always-migrating households.

In the entire sample of young women, the maximum value of λ can take – to ensure that p_{AN} and p_{CC} are not negative – is 1.7, which is the minimum of $\frac{P(M=1,D=0|Z=0)}{P(D=1|Z=0)}$ and $\frac{P(D=1|Z=1)}{P(D=1|Z=0)} - 1$. To investigate the sensitivity of the bounds with respect λ, I plot in figure 2 the the variation of the bounds on θ_{CN} with the value of λ. I also derive non-parametric bounds using the the three different population categories based on the ex-ante propensity to migrate $p(X)$. I assume that the parameter λ is identical across the three categories of low, medium and high pscore $p(X)$. As a result, the range of λ for which the bounds on θ_{CN} are defined across the entire sample is bounded from above by the minimum of $\min(\frac{P(M=1,D=0|Z=0)}{P(D=1|Z=0)}, \frac{P(D=1|Z=1)}{P(D=1|Z=0)} - 1)$ across the three categories of $p(X)$. The maximum value λ can take to ensure that existence of bounds across all categories of $p(X)$ is 0.7.

With respect to participation in self-employed activity, it is apparent in figure 2 that the lower bound on θ_{CN} using the three categories based the index $p(X)$ is around 10%, higher than zero. This suggests that irrespective of the value of λ , the causal impact of migration θ_{CN} seems to be positive and higher

than 10%. With respect to participation in non-rural jobs, the upper bound on θ_{CN} which uses the index $p(X)$ seems to be negative, around -5%, for all possible values of λ. This indicates that the causal impact of migration θ_{CN} is to reduce participation in the non-agricultural labor market among young women staying in Mexico. Overall, the result that migration triggers a re-allocation of labor among left-behinds appears to be robust to the violation of assumption 6.

Table 5: Bounds on θ_{CN} - No covariates used

	Self-employed labor	Non agricultural wage labor
LATE	$0,276^{**}$	$-0,239^{**}$
	(0,111)	(0,121)
$\widehat{\theta_{CN}}^{IV}$	$0,191^{**}$	-0,161 *
	(0,089)	(0,090)
bounds on θ_{CN}	[-0,078; 0,399]	[-0,603 ; 0,222]
p_{AA}	0,021	0,021
p_{AN}	0,044	0,044
p_{NN}	0,692	0,692
p_{CC}	0,079	0,079
p_{CN}	0,164	0,164
$\mathbb{E}[Y(1,0)\|CN]$	0,026	-0,033
$\mathbb{E}^{U}[Y(0,0)\|CN]$	0,105	0,570
$\mathbb{E}^{L}[Y(0,0)\|CN]$	-0,373	-0,255
N	1521	1521

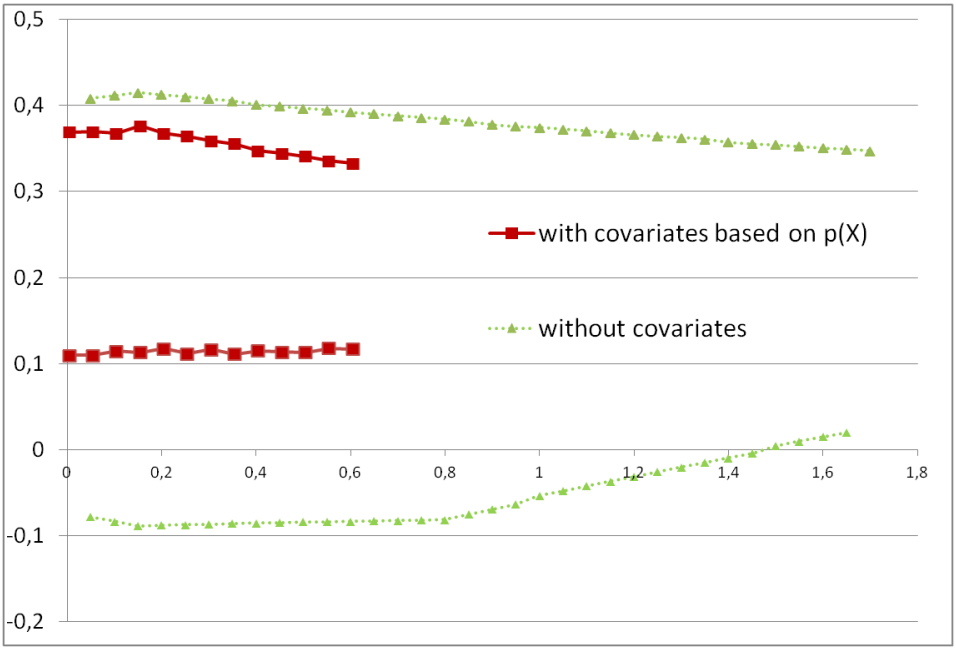
$\widehat{\theta_{CN}}^{IV}$ is the corrected IV estimator under the assumption of the absence of intra-household selection $\mathbb{E}[Y(0,0)|CN] = \mathbb{E}[Y(0,0)|CC]$. Standard errors in parentheses from 99 bootstrap replications. * denotes that estimates are statistically different from zero at the 10% level, ** at the 5% level.

Table 6: Bounds on θ_{CN} . Migration propensity $p(X)$ used to define three categories with high, medium and low propensity

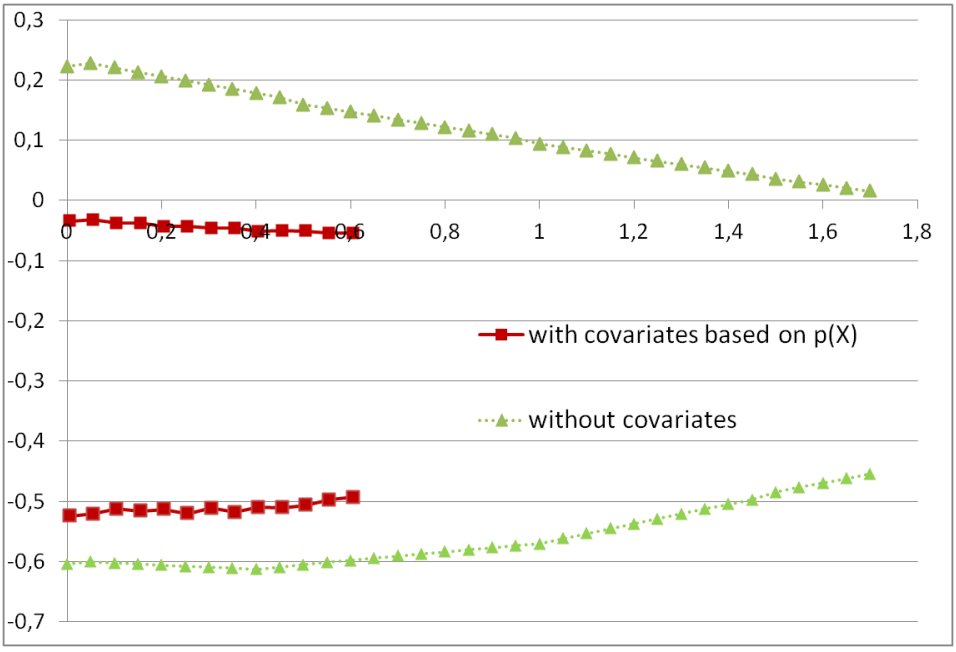
	Self-employed labor	Non agricultural wage labor
LATE	$0,316^{**}$	$-0,321^{**}$
	(0,124)	(0,136)
$\widehat{\theta_{CN}}^{IV}$	0,249**	-0,308 **
	(0,116)	(0,113)
bounds on θ_{CN}	[0,247 ; 0,377]	[-0,545 ;-0,084]
N	1521	1521

Figure 2: Sensitivity of Bounds for θ_{CN} with respect to λ : (i) without covariates and (ii) using three categories of $p(X)$ with high, medium and low propensity to migrate

(a) Self-employed labor



(b) Non agricultural wage labor



8 Empirical application: migration and children education

Another strand of research has investigated the effect of migration on educational attainment of children staying behind. Using the rural sample of 12 to 18 years old children in the MxFLS survey, I revisit previous estimates. As table 7 describes, school attendance drops from 95% at age 12 to 71% at age 15 and 46% at age 16. About 18% of children live in household who will participate in migration by 2005. About 10% of children older than 14 will migrate themselves to the U.S. within the next 3 years of the baseline survey (2002).

I estimate the impact of migration on the variation in school attendance between 2002 and 2005. I use the same instrument as previously, i.e. the past emigration rate by municipality. I derive non parametric bounds under the assumption that the instrument does not affect the intra-household selection of migrants (assumption 6). I also derive bounds using three categories of individual propensity to migrate estimated with a probit model using initial demographic characteristics of the individuals and of the households. As table 8 shows, the standard LATE estimator indicates that migration reduces attendance by 40% . The corrected IV estimator $\widehat{\theta_{CN}^{IV}}$ suggests that the LATE suffers from downward bias, even in the absence of intra-household selection of migrant(s). In the presence of such selection, the bounds do not exclude that the causal impact may be null. This suggests that children might be positively selected into migration among households engaging into migration (i.e. migrating children would have stayed longer in school relative to children left behind).

Table 7: Children sample in the MxFLS survey

age	school attendance		Migration 2005-2002	
	in 2002	2005-2002	$M=1$	$D=1$
12	0,95	-0,38	0,14	0,03
13	0,90	-0,43	0,16	0,04
14	0,84	-0,42	0,17	0,09
15	0,71	-0,43	0,17	0,09
16	0,46	-0,28	0,19	0,12
17	0,42	-0,25	0,17	0,09
18	0,26	-0,15	0,20	0,13
Observations	1972	1972	1972	1972

Table 8: Bounds on causal impact on migration on school attendance (under assumption 6)

	No covariates	Covariates based on p(x)
LATE	$-0,42^{**}$	$-0,43^{**}$
	(0,15)	(0,15)
$\widehat{\theta_{CN}}^{IV}$	-0,25 **	-0,24 **
	(0,09)	(0,09)
bounds	[-0,50 ;-0,001]	[-0,43; -0,007]
N	1972	1972

$\widehat{\theta_{CN}}^{IV}$ is the corrected IV estimator under the assumption of the absence of intra-household selection $\mathbb{E}[Y(0,0)|CN] = \mathbb{E}[Y(0,0)|CC]$. Standard errors in parentheses from 99 bootstrap replications. * denotes that estimates are statistically different from zero at the 10% level, ** at the 5% level.

9 Conclusion

This paper examines the identification of the causal effect of migration on the family left behind in presence of double-selection. The first selection problem arises from the well-studied self-selection of households into migration and the endogenous family decision as to whether send or not a migrant. The second selection problem arises from the selection of migrant individuals within households and the endogenous decision as to which members migrate and which stay behind at origin. The complex identification problem this second form of selectivity generates has largely been ignored in the literature.

Similarly to Steinmayr (2014) , I use principal stratification to model migration decisions and structure the identification problem. This allows deriving non parametric bounds on the causal effect of migration under different sets of assumptions in a setting with double-selection. Using panel data drawn from the Mexican Family Life Survey , I illustrate the approach by estimating bounds on the migration impact on two outcome: the labor supply of young women left behind and the school attendance of children remaining in Mexico. Results suggest that ignoring the intra-household selection problem may lead to biased estimates, usually overstating the true magnitude of the causal effect.

Most importantly, this paper shows that point-identification of the causal effect of migration requires strong, if not unrealistic, assumptions which are rarely met in practice even with ideal experimental data. Empirical applications indicate that a combination of weaker assumptions can instead provide sufficient identifying power to derive informative bounds. Partial identification appears therefore as a practical, judicious and promising solution for migration research.

References

Acosta, P. (2006). Labor supply, school attendance, and remittances from international migration: the case of El Salvador. *World Bank Policy Research Working Paper*.

Adams, R. H. (2011). Evaluating the Economic Impact of International Remittances On Developing Countries Using Household Surveys: A Literature Review. *Journal of Development Studies 47*(6), 809–828.

Amuedo-Dorantes, C. and S. Pozo (2006). Migration, remittances, and male and female employment patterns. *The American economic review 96*(2), 222–226.

Antman, F. (2013). The impact of migration on family left behind. In K. F. Zimmermann and A. Constant (Eds.), *International Handbook on the Economics of Migration*. Cheltenham, UK: Edward Elgar Publishing Limited.

Antman, F. M. (2012). Elderly Care and Intrafamily Resource Allocation when Children Migrate. *Journal of Human Resources 47*(2), 331–363.

Binzel, C. and R. Assaad (2011). Egyptian men working abroad: Labour supply responses by the women left behind. *Labour Economics 18*, S98–S114.

Chen, X. and C. Flores (2012). Bounds on treatment effects in the presence of sample selection and noncompliance: the wage effects of Job Corps. *Mimeo* (October).

Chort, I. and J.-n. Senne (2013). Intra-household Selection into Migration : Evidence from a Matched Sample of Migrants and Origin Households in Senegal. *WORKING PAPER N° 2013 –35, Paris School of Economics 33*(0).

De Janvry, A. and E. Sadoulet (2001). Income strategies among rural households in Mexico: The role of off-farm activities. *World development 29*(3).

Frangakis, C. E. and D. B. Rubin (2002). Principal stratification in causal inference. *Biometrics 58*(1), 21–29.

Frolich, M. (2007). Nonparametric IV estimation of local average treatment effects with covariates. *Journal of Econometrics 139*(1), 35–75.

Gibson, J., D. McKenzie, and S. Stillman (2011). The impacts of international migration on remaining household members: omnibus results from a migration lottery program. *Review of Economics and Statistics* (202).

Hanson, G. (2007). Emigration, remittances, and labor force participation in Mexico. *Integration and Trade Journal 27*(2005), 73–105.

Heckman, J. (1974). Shadow prices, market wages, and labor supply. *Econometrica 42*(4), 679–694.

Hernandez-Leon, R. (2008). *Metropolitan migrants: the migration of urban Mexicans to the United States*. University of California Press.

Huber, M. and G. Mellace (2013). Sharp IV bounds on average treatment effects under endogeneity and noncompliance. *Mimeo*.

Ibarraran, P. and D. Lubotsky (2007). Mexican Immigration and Self-Selection: New Evidence from the 2000 Mexican Census. In G. J. Borjas (Ed.), *Mexican Immigration to the United States*, Number May, Chapter 5, pp. 159–192. University of Chicago Press.

Imbens, G. W. and J. D. Angrist (1994). Identification and Estimation of Local Average Treatment Effects. *Econometrica 62*(2), 467–75.

Kitagawa, T. (2009). Identification region of the potential outcome distributions under instrument independence. *cemmap working paper, No. CWP30/09*.

Lee, D. S. (2009). Training , Wages , and Sample Selection : Estimating Sharp Treatment. *The Review of Economic Studies 76*(3), 1071–1102.

Lokshin, M. and E. Glinskayai (2008). The effect of male migration for work on employment patterns of females in Nepal. *World Bank Policy Research Working Paper* (October).

Lucas, R. E. (1997). Internal migration in developing countries. In M. Rosenzweig and O. Stark (Eds.), *Handbook of population and family economics, vol. 1B.*, Chapter 13, pp. 721–798. Amsterdam: Elsevier Science B.V.

McKenzie, D. and H. Rapoport (2007). Network effects and the dynamics of migration and inequality: Theory and evidence from Mexico. *Journal of Development Economics 84*(1), 1–24.

McKenzie, D. and H. Rapoport (2011). Can migration reduce educational attainment? Evidence from Mexico. *Journal of Population Economics 24*(4), 1331–1358.

McKenzie, D. and D. Yang (2010). Experimental Approaches in Migration Studies. *World Bank Policy Research Working Paper 5395*.

Mendola, M. and C. Carletto (2012). Migration and gender differences in the home labour market: Evidence from Albania. *Labour Economics 19*(6), 870–880.

Mu, R. and D. van de Walle (2011). Left behind to farm? Women's labor re-allocation in rural China. *Labour Economics 18*, S83–S97.

Murard, E. (2013). Family left behind, labor supply and household production : Theory and evidence from Mexican migration. *Paris School of Economics*.

Rodriguez, E. and E. Tiongson (2001). Temporary migration overseas and household labor supply: evidence from urban Philippines. *International Migration Review 35*(3), 709–725.

Rubalcava, L. and G. Teruel (2008). User's Guide for the Mexican Family Life Survey Second Wave.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology 66*(5), 688–701.

Stark, O. (1991). *The migration of labor*. Oxford: Blackwell.

Stark, O. and D. E. Bloom (1985). The new economics of labor migration. *The american Economic review*, 173–178.

Stark, O. and R. Lucas (1988). Migration, remittances, and the family. *Economic Development and Cultural Change 36*(3), 465–481.

Steinmayr, A. (2014). When a Random Sample is Not Random. Bounds on the Effect of Migration on Children Left Behind. *Swiss Institute for Empirical Economic Research, U*, Swiss Institute for Empirical Economic Research, U.

Teruel, G., E. Arenas, and L. Rubalcava (2012). Migration in the Mexican Family Life Survey. In *Migration and Remittances: Trends, Impacts and New Challenges* (Rowman and ed.).

Woodruff, C. and R. Zenteno (2007). Migration networks and microenterprises in Mexico. *Journal of Development Economics 82*(2), 509–528.

Zhang, J. L. and D. B. Rubin (2003). Estimation of Causal Effects via Principal Stratification When Some Outcomes are Truncated by "Death". *Journal of Educational and Behavioral Statistics 28*(4), 353–368.

10 Appendix

Appendix A1 : additional tables and figures

Figure 3: Cut-off for binary instrument . 1995-2000 Emigration rate to the U.S. by municipality (2000 Mexican census)

(a)

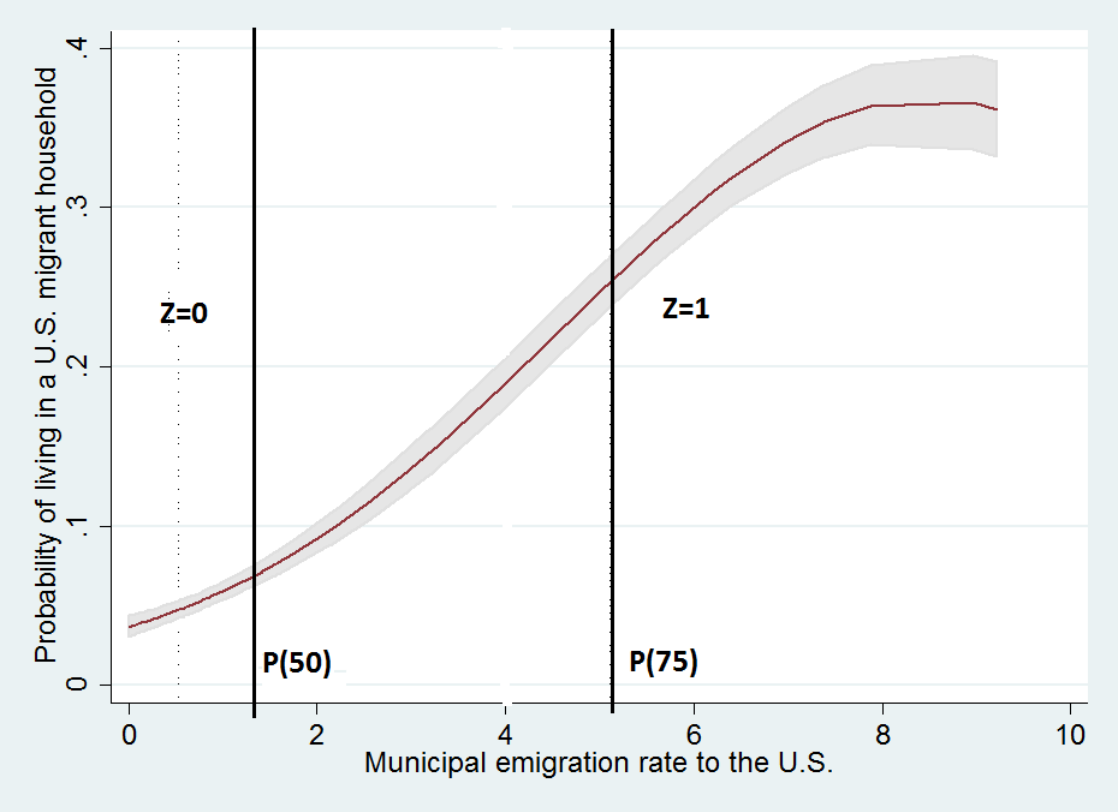


Table 9: Latent strata with and without assumptions

				Latent group				
				(1)	(2)	(3)	(4)	(5)
$M(1)$	$M(0)$	$D(1)$	$D(0)$	All	$M(z)=0 \Rightarrow D(z)=0$	Assum. 4	Assum. 6	Assum. 7 instead of 6
1	1	1	1	AA	AA	AA	AA	AA
1	1	1	0	AC	AC	AC		AC
1	1	0	1	AD	AD	AD		
1	1	0	0	AN	AN	AN	AN	AN
1	0	1	1	CA				
1	0	1	0	CC	CC	CC	CC	CC
1	0	0	1	CD				
1	0	0	0	CN	CN	CN	CN	CN
0	1	1	1	DA				
0	1	1	0	DC				
0	1	0	1	DD	DD			
0	1	0	0	DN	DN			
0	0	1	1	NA				
0	0	1	0	NC				
0	0	0	1	ND				
0	0	0	0	NN	NN	NN	NN	NN

Column (1) shows all 16 strata.
Column (2) shows remaining strata after the implication $M(z)=0 \Rightarrow D(z)=0$
Column (3) shows remaining strata after Assumption 4 has been made
Column (3) shows remaining strata after Assumption 6
Column (3) shows remaining strata after Assumption 6 is replaced with Assumption 7

Table 10: Initial (2002) participation in the labor force among young women living in U.S. migrant-sending households : differences between migrant and left behind women .

	(1)	(2)	(3)
Participation in :			
Any work	0.149**	0.189***	0.250*
	(0.066)	(0.066)	(0.127)
Non rural jobs	0.131**	0.148**	0.185
	(0.059)	(0.060)	(0.120)
Self-employed work	0.011	0.029	0.038
	(0.041)	(0.041)	(0.054)
Controls:			
Household characteristics†		✓	
Household fixed effets			✓
N	229	229	229

Each number corresponds to a different regression .
Standard errors in (). Significance levels: * $p<0.10$, ** $p<0.05$, *** $p<0.01$
† : individual age and education, household size, number of elderly, number of children under 12 , highest education in the family, initial social and private transfers received, and initial wealth of the household

Table 11: Probability individual migration among young women in rural area

	(1)	
	Migration U.S.	
	b	se
nbmaleage1435	0.006	(0.076)
nbmaleage3654	-0.035	(0.137)
nbfemaleage3654	0.269*	(0.150)
nbage13minus	0.059	(0.039)
nb_60	0.243**	(0.101)
ln_socialnonlabincome	0.002	(0.017)
ln_HH_transfer02	0.027*	(0.016)
hh_higestgrade3	-0.153	(0.207)
hh_higestgrade4	-0.620*	(0.328)
wealthindex	0.105**	(0.051)
wealthindexsq	-0.030	(0.021)
ln_pop95	-0.052	(0.062)
ln_ave_dis	0.120**	(0.057)
ln_medincome_mun2000	-0.295**	(0.120)
age	0.141	(0.111)
age2	-0.004*	(0.002)
educ primary	-0.105	(0.189)
educ secondary	0.206	(0.238)
educ post secondary	0.278	(0.402)
spouse not resident	0.742***	(0.252)
divorced	0.329	(0.378)
single never married	0.437**	(0.191)
_cons	-1.535	(1.589)
N	1614	

Table 12: LATE and bounds within each cell defined by Pscore

Self-employed labor

Category pscore	LATE	Lower bound	Upper bound	p_{CN}	N
high pscore	0,248	-0,068	0,345	0,154	538
medium pscore	0,230	0,230	0,249	0,147	538
low pscore	0,558	0,558	0,558	0,143	538

Non agricultural wage labor

Category pscore	LATE	Lower bound	Upper bound	p_{CN}	N
high pscore	-0,334	-0,846	0,324	0,154	538
medium pscore	-0,336	-0,336	-0,320	0,147	538
low pscore	-0,240	-0,240	-0,240	0,143	538

Appendix A2 : Derivation lower and upper bounds $\mathbb{E}^L[Y(0,0)|CN]$ and $\mathbb{E}^U[Y(0,0)|CN]$ under assumptions 2 to 6

The observed outcome for the group $\{Z=0,M=0,D=0\}$ is a mixture of the outcomes of the three strata CN, CC and NN and the outcome of the stratum NN is point identified. Indeed, $Y(0,0)$ for the latent group NN is directly observed in the observed group $\{Z=1,M=,0,D=0\}$:

$$\mathbb{E}[Y(0,0)|NN] = \overline{Y}^{100}$$

I introduce additional notation to describe the bounds. Let y_r^{zmd} be the r-th quantile of Y in the observed group $\{Z=z,M=m,D=d\}$ and let the mean outcome in this cell for the outcomes between y_r^{zmd} and $y_{r'}^{zmd}$ be

$$\overline{Y}(y_r^{zmd} \leq Y \leq y_{r'}^{zmd}) = \mathbb{E}[Y|Z=z,M=m,D=d,y_r^{zmd} \leq Y \leq y_{r'}^{zmd}]$$

Denote also $\alpha_{CN} = \frac{p_{CN}}{p_{CN}+p_{NN}+p_{CC}}$, $\alpha_{NN} = \frac{p_{NN}}{p_{CN}+p_{NN}+p_{CC}}$ and $\alpha_{CC} = 1-\alpha_{CN}-\alpha_{NN}$ the conditional probabilities in the observed group $\{Z=0,M=0,D=0\}$. The idea behind the bounds proposed by Chen and Flores (2012) is to calculate the lowest and highest possible values of $\mathbb{E}[Y(0,0)|CN]$ that are consistent withe the constraint that $\mathbb{E}[Y(0,0)|NN] = \overline{Y}^{100}$. I now describe their procedure to derive the lower bound.

To begin, consider the problem without the constraint and ignore the information about NN. In this case, I can directly apply the trimming procedure of Zhang and Rubin(2003) and Lee(2009) , just as described in the previous section. $\mathbb{E}[Y(0,0)|CN]$ can be bounded from below by the expected value of Y for the α_{CN} fraction of the *smallest* values of Y in the cell $\{Z=0,M=0,D=0\}$, that is $\overline{Y}(Y \leq y^{000}_{\alpha_{CN}})$. Next, I check whether this solution is consistent with constraint that $\mathbb{E}[Y(0,0)|NN] = \overline{Y}^{100}$. To do this,I construct the "wort-case" scenario lower bound for $\mathbb{E}[Y(0,0)|NN]$ by assuming that all observations that belong to the NN latent group are at the bottom of the remaining observations in the cell $\{Z=0,M=0,D=0\}$. This yields $\overline{Y}(y^{000}_{\alpha_{CN}} \leq Y \leq y^{000}_{\alpha_{CN}+\alpha_{NN}})$. If $\overline{Y}^{100} \geq \overline{Y}(y^{000}_{\alpha_{CN}} \leq Y \leq y^{000}_{\alpha_{CN}+\alpha_{NN}})$, the unconstrained solution is consistent with the constraint and the lower bound for $\mathbb{E}[Y(0,0)|CN]$ is $\overline{Y}(Y \leq y^{000}_{\alpha_{CN}})$ – similar to Lee's bound. If the constraint is not satisfied, I construct the "worst-case" scenario lower bound for $\mathbb{E}[Y(0,0)|CN]$ by placing all the observations NN and CN at the bottom of the distribution of Y [24] . Thus, the lower bound $\mathbb{E}^L[Y(0,0)|CN]$ can be derived from the equation :

$$\overline{Y}(Y \leq y^{000}_{\alpha_{CN}+\alpha_{NN}}) = \frac{\alpha_{CN}}{\alpha_{CN}+\alpha_{NN}} * \mathbb{E}^L[Y(0,0)|CN] + \frac{\alpha_{NN}}{\alpha_{CN}+\alpha_{NN}} * \overline{Y}^{100}$$

The upper bound is derived in a similar way as the lower bound, but now by placing the observations in the corresponding strata in the upper part of the distribution of Y in the cell $\{Z=0,M=0,D=0\}$. It follows that the lower bound is (Chen and Flores (2012)) :

[24] Intuitively,the fact that $\overline{Y}^{100} < \overline{Y}(y^{000}_{\alpha_{CN}} \leq Y \leq y^{000}_{\alpha_{CN}+\alpha_{NN}})$ implies that some observations in the NN stratum must at the bottom α_{CN}fraction of the smallest values of Y. Thus, $\overline{Y}(Y \leq y^{000}_{\alpha_{CN}})$ is not a sharp lower bound for $\mathbb{E}[Y(0,0)|CN]$.

$$\mathbb{E}^L[Y(0,0)|CN] = \begin{cases} \overline{Y}(Y \leq y^{000}_{\alpha_{CN}}) \text{ if } \overline{Y}(y^{000}_{\alpha_{CN}} \leq Y \leq y^{000}_{\alpha_{CN}+\alpha_{NN}}) \leq \overline{Y}^{100} \\ \dfrac{\alpha_{CN}+\alpha_{NN}}{\alpha_{CN}} * \overline{Y}(Y \leq y^{000}_{\alpha_{CN}+\alpha_{NN}}) - \dfrac{\alpha_{NN}}{\alpha_{CN}} * \overline{Y}^{100} \text{ otherwise} \end{cases} \tag{7}$$

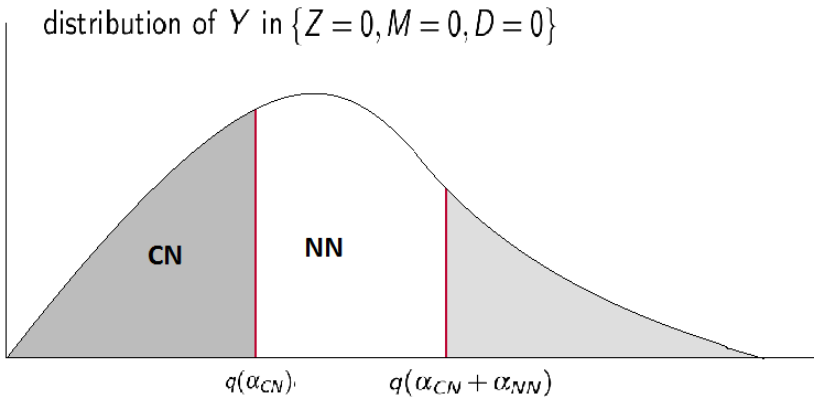
and the upper bound:

$$\mathbb{E}^U[Y(0,0)|CN] = \begin{cases} \overline{Y}(Y \geq y^{000}_{1-\alpha_{CN}}) \text{ if } \overline{Y}(y^{000}_{1-\alpha_{CN}-\alpha_{NN}} \leq Y \leq y^{000}_{1-\alpha_{CN}}) \geq \overline{Y}^{100} \\ \dfrac{\alpha_{CN}+\alpha_{NN}}{\alpha_{CN}} * \overline{Y}(Y \leq y^{000}_{1-\alpha_{CN}-\alpha_{NN}}) - \dfrac{\alpha_{NN}}{\alpha_{CN}} * \overline{Y}^{100} \text{ otherwise} \end{cases} \tag{8}$$

Bounds for the causal effect can be constructed by combining these bounds with the point identified potential outcome under treatment for the latent group CN:

$$\begin{aligned} \theta^U &= \mathbb{E}[Y(1,0)|CN] - \mathbb{E}^L[Y(0,0)|CN] \\ \theta^L &= \mathbb{E}[Y(1,0)|CN] - \mathbb{E}^U[Y(0,0)|CN] \end{aligned} \tag{9}$$

Figure 4: Unconstrained lower bound $\mathbb{E}^L[Y(0,0)|CN]$



Appendix A3 : Alternative bounds under assumption 7 instead of 6

I thus drop assumption 6 in this subsection. Now the existence of the latent strata AC and AD cannot be ruled out anymore. I then make the alternative assumption that the effect of the instrument on *individual* migration is monotonic: every individual is more likely to migrate if $Z = 1$ than if $Z = 0$. This excludes the existence of defiers and therefore the existence of the stratum AD [25]

Assumption 7. *Monotonicity of D in Z*

$$D(1,1) \geq D(1,0)$$

The proportion of the latent groups are no more identified.Indeed, there are five unknown strata proportions for only four known probabilities (linear independent equations) [26]. I therefore introduce an additional parameter λ, which equals the ratio of the share of always migrant individuals to the the share of complier individuals in always migrant households :

$$p_{AC} = \lambda p_{AA}$$

We know that $p_{AA} = P(D = 1|Z = 0)$ that $P(M = 1, D = 0|Z = 0) = p_{AN} + \lambda p_{AA}$ and also that $P(D = 1|Z = 1) = p_{CC} + (1 + \lambda)p_{AA}$[27]. These three equations and the requirement that $p_{AN} \geq 0$ and $p_{CC} \geq 0$ give a interval of possible values of λ, i.e.

$$0 \leq \lambda \leq \lambda_{max} = \min(\lambda^{an}_{max}, \lambda^{cc}_{max})$$

$$\text{with } \lambda^{an}_{max} = \frac{P(M = 1, D = 0|Z = 0)}{P(D = 1|Z = 0)} \text{ and } \lambda^{cc}_{max} = \frac{P(D = 1|Z = 1)}{P(D = 1|Z = 0)} - 1$$

When $\lambda = 0$, we are back to the previous situation where the share of complier individuals in always migrating households AC is zero. If $\lambda^{an}_{max} < \lambda^{cc}_{max}$ then $p_{AN} = 0$ when λ reaches his upper bound : there are no never-migrants in always migrant-sending households. This means that this type of households (A) migrate as a whole, with the entire family, when $Z = 1$. In this case the outcome under treatment

[25] As the rest of the paper will show, this assumption only affects the bounds for $\mathbb{E}[Y(1,0)|CN]$: they are less tight if the existence of defiers is not rule out

[26] The 6th strata proportions is one minus the sum of the others. There are only 4 independent equations because

$$P(M = 1, D = 1|Z) + P(M = 1, D = 0|Z) + P(M = 0, D = 0|Z) = 1$$

[27] The share of the latent groups can be expressed as :

$$\begin{aligned}
p_{AA} &= P(M = 1, D = 1|Z = 0) \\
p_{AC} &= \lambda p_{AA} \\
p_{NN} &= P(M = 0, D = 0|Z = 1) \\
p_{AN} &= P(M = 1, D = 0|Z = 0) - p_{AC} \\
p_{CN} &= P(M = 1, D = 0|Z = 1) - P(M = 1, D = 0|Z = 0) + p_{AC} \\
p_{CC} &= P(M = 0, D = 0|Z = 0) - p_{CN} - p_{NN}
\end{aligned}$$

$\mathbb{E}[Y(1,0)|CN]$ is identified. If $\lambda_{max}^{an} > \lambda_{max}^{cc}$ then $p_{CC} = 0$ when λ reaches λ_max: in complier households, the family members under investigation never migrate themselves. IN this case the outcome under control $\mathbb{E}[Y(0,0)|CN]$ is identified [28]. These two symmetric situations will be considered in the empirical application.

Under Assumptions 7 (instead of 6) it is no longer possible to point-identify $\mathbb{E}[Y(1,0)|CN]$, expect when $\lambda = 0$ or when $\lambda = \lambda_{max} = \lambda_{max}^{an}$ [29]. For λ strictly within its interval, it is however possible to derive sharp bounds for $\mathbb{E}[Y(1,0)|AN]$ and in further consequence for $\mathbb{E}[Y(1,0)|CN]$ [30]. Indeed the potential outcome under treatment $Y(1,0)$ for the latent group CN is observed as part of the mixture distribution in the observed group $\{Z = 1, M = 1, D = 0\}$:

$$\overline{Y}^{110} = \frac{p_{CN}\mathbb{E}[Y(1,0)|CN] + p_{AN}\mathbb{E}[Y(1,0)|AN]}{p_{CN} + p_{AN}}$$

It follows that sharp bounds for $\mathbb{E}[Y(1,0)|AN]$ willyield sharp bounds for $\mathbb{E}[Y(1,0)|CN]$

I use the procedure of Huber and Mellace(2013) to derive sharp bounds for $\mathbb{E}[Y(1,0)|AN]$. First let $\alpha_{AN}^{110} = \frac{p_{AN}}{p_{AN}+p_{CN}}$ denote the fraction of AN in the group $\{Z = 1, M = 1, D = 0\}$ and $\alpha_{AN}^{010} = \frac{p_{AN}}{p_{AN}+p_{AC}}$ denote the fraction of AN is the group $\{Z = 0, M = 1, D = 0\}$. The conditional distribution $Y(1,0)|AN$ is observed in each of these the two groups . Within each of this cell, I can bound $\mathbb{E}[Y(1,0)|AN]$ from below by the expected value of Y in the α_{AN}^{z10} fraction of the smallest value of Y for $z = 1,0$. The sharp lower for $\mathbb{E}[Y(1,0)|AN]$ is the maximum of the two. For the upper bound, I take the α_{AN}^{z10} fraction of the largest value of Y and then the minimum of the two. This yields the upper and lower bounds for $\mathbb{E}[Y(1,0)|CN]$:

$$\begin{aligned}
\mathbb{E}^L[Y(1,0)|CN] &= \frac{p_{CN} + p_{AN}}{p_{CN}} * \overline{Y}^{110} - \frac{p_{AN}}{p_{CN}} * \min\left\{\overline{Y}(Y \geq y^{110}_{1-\alpha_{AN}^{110}}), \overline{Y} \geq y^{010}_{1-\alpha_{AN}^{010}})\right\} \\
\mathbb{E}^U[Y(1,0)|CN] &= \frac{p_{CN} + p_{AN}}{p_{CN}} * \overline{Y}^{110} - \frac{p_{AN}}{p_{CN}} * \max\left\{\overline{Y}(Y \leq y^{110}_{\alpha_{AN}^{110}}), \overline{Y} \leq y^{010}_{\alpha_{AN}^{010}})\right\}
\end{aligned} \tag{10}$$

Bounds for the causal effect can be constructed by combining the bounds for the potential outcome of stratum CN under control with the bounds for potential outcome of CN under treatment :

$$\begin{aligned}
\theta^U &= \mathbb{E}^U[Y(1,0)|CN] - \mathbb{E}^L[Y(0,0)|N] \\
\theta^L &= \mathbb{E}^L[Y(1,0)|CN] - \mathbb{E}^U[Y(0,0)|N]
\end{aligned}$$

[28] When $\lambda = \lambda_{max} = \lambda_{max}^{cc}$:

$$\mathbb{E}[Y(0,0)|CN] = \frac{P(M=0|Z=0) * Y^{000} - P(M=0|Z=1) * Y^{100}}{P(M=1|Z=1) - P(M=1|Z=0)}$$

[29] In which cases, and noting $P(m,d|z) = P(M = m, D = d|Z = z)$ to shorten notation :

When $\lambda = 0$: $\mathbb{E}[Y(1,0)|CN] = \frac{P(1,0|1)\overline{Y}^{110} - P(1,0|0)\overline{Y}^{010}}{P(1,0|1) - P(1,0|0)}$ - see eq (3)

When $\lambda = \lambda_{max} = \lambda_{max}^{an}$: $\mathbb{E}[Y(1,0)|CN] = \overline{Y}^{110}$

[30] See Huber and Mellace (2013) for the proof of sharpness of these bounds.

Finally, note that the bounds θ^U and θ^L are not monotonic functions of λ in general. This is because the bounds for $\mathbb{E}[Y(1,0)|CN]$ are not monotonic with respect to λ [31] .Therefore the sensitivity of the bounds forθ_{CN} with respect to λ remains an empirical question. However, it can be shown that the bounds for $\mathbb{E}[Y(0,0)|CN]$ unambiguously contract with λ. This is because p_{CC} decreases with λ : the share of the group CC for which the potential outcome under control $Y(0,0)$ is unknown shrinks as λ augments. Therefore in the observed group $\{M=0, D=0, Z=0\}$ the mixture of the distribution of the latent groups CC + CN + NN become closer to the mixture CN+ NN. Since the expected value of $Y(0,0)$ for NN is point identified already, the potential outcome for CN can be inferred with more precision. In others words, the bounds are tighter.

Table 13: Latent and observed groups without exclusion restriction of Z with respect to D

$M(1)$	$M(0)$	$D(1)$	$D(0)$	Latent group
1	1	1	1	AA
1	1	1	0	AC
1	1	0	0	AN
1	0	1	0	CC
1	0	0	0	CN
0	0	0	0	NN

		D=0	D=1
$Z=1$	$M=1$	**AN, CN**	AA ,AC ,CC
	$M=0$	**NN**	
$Z=0$	$M=1$	**AC, AN**	AA
	$M=0$	**CC, CN ,NN**	

[31] To see this, note than p_{AN} is decreasing with λ and p_{CN} is increasing. Therefore the ratio $\frac{p_{AN}}{p_{CN}}$ is decreasing . Note also that α_{AN}^{110} and α_{AN}^{100} are both decreasing. To fix idea let assume that $Y>0$. Then it is clear that $\forall z\ \overline{Y}(Y \leq y_{\alpha_{AN}^{z10}}^{z10})$ are decreasingfunction of λ since a smaller fraction of the smallest value o Y are averaged out. So the max of the two is also decreasing . The firstterm of $\mathbb{E}^U[Y(1,0)|CN]$ is decreasing as well because $\frac{p_{CN}+p_{AN}}{p_{CN}}$ is decreasing. So $\mathbb{E}^U[Y(1,0)|CN]$ is equal to adecreasing term minus another decreasing term. IN consequence, it is a non monotonic function of λ.

Appendix A4 : bias in the LATE

Computation : what does LATE estimate ?

$$\begin{aligned} LATE &= \equiv \frac{\mathbb{E}[Y|Z=1,D=0]-\mathbb{E}[Y|Z=0,D=0]}{\mathbb{E}[M|Z=1,D=0]-\mathbb{E}[M|Z=0,D=0]} \\ &= \frac{DY}{DP} \end{aligned} \tag{11}$$

Note $P(m|d,z) = P(M=m|D=d,Z=z)$ and $\overline{Y}^{zmd} = E[Y|Z=z,M=m,D=d]$

$$\overline{Y}^{110} = \frac{p_{CN}}{p_{CN}+p_{AN}}E[Y(1,0)|CN] + \frac{p_{AN}}{p_{CN}+p_{AN}}E[Y(1,0)|AN]$$
$$\overline{Y}^{100} = E[Y(0,0)|NN]$$
$$\overline{Y}^{010} = E[Y(1,0)|AN]$$
$$\overline{Y}^{000} = \frac{p_{CC}}{p_{CN}+p_{NN}+p_{CC}}E[Y(0,0)|CC] + \frac{p_{CN}}{p_{CN}+p_{NN}+p_{CC}}E[Y(0,0)|CN] + \frac{p_{NN}}{p_{CN}+p_{NN}+p_{CC}}E[Y(0,0)|NN]$$

And, noting $p_{xN} = p_{AN}+p_{CN}+p_{NN}$:

$$P(m=1|d=0,z=1) = \frac{p_{CN}+p_{AN}}{p_{xN}}$$
$$P(m=1|d=0,z=0) = \frac{p_{AN}}{p_{xN}+p_{CC}}$$

Then :

$$DY = P(1|0,1)\overline{Y}^{110} + (1-P(1|0,1))\overline{Y}^{100} - P(1|0,0)\overline{Y}^{010} - (1-P(1|0,0))\overline{Y}^{000}$$
$$\begin{aligned} DY = &\frac{p_{CN}}{p_{xN}}E[Y(1,0)|CN] + \frac{p_{AN}}{p_{xN}}E[Y(1,0)|AN] + \frac{p_{NN}}{p_{xN}}E[Y(0,0)|NN] \\ &- \frac{p_{AN}}{p_{xN}+p_{CC}}E[Y(1,0)|AN] - \frac{p_{CC}}{p_{xN}+p_{CC}}E[Y(0,0)|CC] - \frac{p_{CN}}{p_{xN}+p_{CC}}E[Y(0,0)|CN] - \frac{p_{NN}}{p_{xN}+p_{CC}}E[Y(0,0)|NN] \end{aligned}$$

$$DP = \frac{p_{CN}+p_{AN}}{p_{xN}} - \frac{p_{AN}}{p_{xN}+p_{CC}} = \frac{p_{AN}p_{CC}+p_{CN}(p_{xN}+p_{CC})}{p_{xN}(p_{xN}+p_{CC})} = \frac{\phi}{p_{xN}(p_{xN}+p_{CC})}$$

SO :

$$\begin{aligned} LATE = \frac{1}{\phi} * [&\underbrace{p_{CN}(p_{xN}+p_{CC})E[Y(1,0)|CN] - p_{CN}p_{xN}E[Y(0,0)|CN]}_{a} + \\ &p_{CC}p_{AN}E[Y(1,0)|AN] + p_{CC}p_{NN}E[Y(0,0)|NN] - p_{CC}p_{xN}E[Y(0,0)|CC]] \end{aligned}$$

writing

$$a = \phi * \theta_{CN} + p_{CC}((p_{AN}+p_{CN})E[Y(0,0)|CN] - p_{AN}E[Y(1,0)|CN])$$

$$LATE = \theta_{CN} + \frac{p_{CC}}{\phi} * [(p_{AN}+p_{CN})E[Y(0)|CN] - p_{AN}E[Y(1)|CN] + p_{AN}E[Y(1)|AN] + p_{NN}E[Y(0)|NN] - p_{xN}E[Y(0)|CC]]$$

$$LATE = \theta_{CN} + \frac{p_{CC}}{\phi} * [$$
$$(p_{AN} + p_{CN})E[Y(0)|CN] - E[Y(0)|CC]) + p_{AN}(E[Y(1)|AN] - E[Y(1)|CN]) + p_{NN}(E[Y(0)|NN] - E[Y(0)|CC])]$$

Computation : Is the LATE within the bounds?

$$LATE = \frac{p_{xN}(P_{xN} + p_{CC})}{\phi} * [\frac{p_{AN} + p_{CN}}{P_{xN}}\overline{Y}^{110} + \frac{p_{NN}}{P_{xN}}\overline{Y}^{100} - \frac{p_{AN}}{P_{xN} + p_{CC}}\overline{Y}^{010} - \frac{p_{CC} + p_{NN} + p_{CN}}{P_{xN} + p_{CC}}\overline{Y}^{000}]$$
$$= \frac{1}{\phi} * [(p_{AN} + p_{CN})(P_{xN} + p_{CC})\overline{Y}^{110} + p_{NN}(p_{xN} + p_{CC})\overline{Y}^{100} - p_{AN}p_{xN}\overline{Y}^{010} - (p_{CC} + p_{NN} + p_{CN})p_{xN}\overline{Y}^{000}$$

$$\theta^U = E[Y(1,0)|CN] - E^L[Y(0,0)|CN] = \frac{p_{AN} + p_{CN}}{P_{CN}}\overline{Y}^{110} - \frac{p_{AN}}{P_{CN}}\overline{Y}^{010} - E^L[Y(0,0)|CN]$$

Since $E^L[Y(0,0)|CN]$ is a function of the distribution of $Y|z = 0, m = 0, d = 0$ and $Y|z = 1, m = 0, d = 0$ then

$$\Delta_u = \theta^U - LATE = Function(p, Y)$$

If there no alway-migrants households : $P_{AA} = p_{AN} = 0$ and $p_{CN} + p_{NN} + p_{CC} = 1$. Then $\phi = p_{CN}$ and

$$LATE = \overline{Y}^{110} + \frac{p_{NN}}{p_{CN}}\overline{Y}^{100} - \frac{p_{NN} + p_{CN}}{p_{CN}}\overline{Y}^{000}$$
$$\theta^U = \overline{Y}^{110} - E^L[Y(0,0)|CN]$$

$$\theta^U - LATE = \frac{p_{NN} + p_{CN}}{p_{CN}}\overline{Y}^{000} - \frac{p_{NN}}{p_{CN}}\overline{Y}^{100} - E^L[Y(0,0)|CN]$$

Recall that

$$\mathbb{E}^L[Y(0,0)|CN] = \begin{cases} \overline{Y}(Y \leq y^{000}_{\alpha_{CN}}) \text{ if } \overline{Y}(y^{000}_{\alpha_{CN}} \leq Y \leq y^{000}_{\alpha_{CN}+\alpha_{NN}}) \leq \overline{Y}^{100} \\ \frac{\alpha_{CN} + \alpha_{NN}}{\alpha_{CN}} * \overline{Y}(Y \leq y^{000}_{\alpha_{CN}+\alpha_{NN}}) - \frac{\alpha_{NN}}{\alpha_{CN}} * \overline{Y}^{100} \text{ otherwise} \end{cases} \quad (12)$$

Note $\varepsilon = \overline{Y}^{100} - \overline{Y}(y^{000}_{\alpha_{CN}} \leq Y \leq y^{000}_{\alpha_{CN}+\alpha_{NN}})$:
If $\varepsilon < 0$, then:

$$\begin{aligned}\theta^U - LATE &= \frac{p_{NN}+p_{CN}}{p_{CN}}[\overline{Y}^{000} - \overline{Y}(Y \leq y^{000}_{\alpha_{CN}+\alpha_{NN}})] \\ &= \frac{p_{CC}}{p_{CN}}[\overline{Y}(Y \geq y_{1-\alpha_{CC}}) - \overline{Y}^{000}] \\ &= \Delta^* > 0\end{aligned}$$

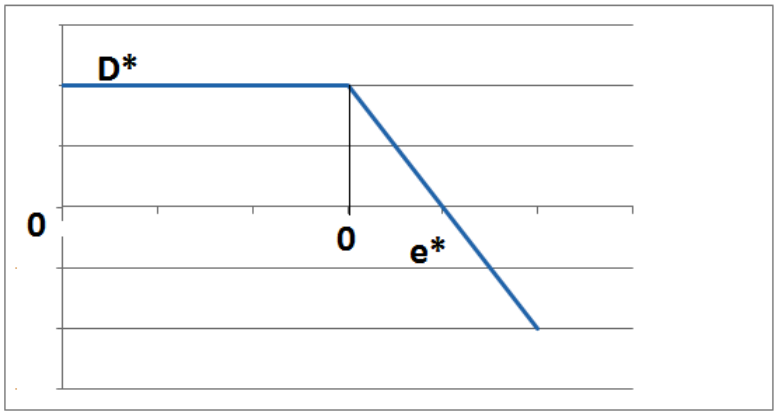
using that: $\overline{Y}^{000} = (p_{NN}+p_{CN})\overline{Y}(Y \leq y^{000}_{\alpha_{CN}+\alpha_{NN}}) + p_{CC}\overline{Y}(Y \geq y_{1-\alpha_{CC}})$

If $\varepsilon > 0$, then:

$$\begin{aligned}\theta^U - LATE &= \frac{p_{CC}}{p_{CN}}[\overline{Y}(Y \geq y_{1-\alpha_{CC}}) - \overline{Y}^{000}] - \frac{p_{NN}}{p_{CN}}\varepsilon \\ &= \Delta^* - \frac{p_{NN}}{p_{CN}}\varepsilon\end{aligned}$$

using that: $\overline{Y}^{000} = p_{CN}\overline{Y}(Y \leq y^{000}_{\alpha_{CN}}) + p_{NN}\overline{Y}(y^{000}_{\alpha_{CN}} \leq Y \leq y^{000}_{\alpha_{CN}+\alpha_{NN}}) + p_{CC}\overline{Y}(Y \geq y_{1-\alpha_{CC}})$

Figure 5: $\theta^U - LATE$ as a function of ε



Note

$$\varepsilon^* = \frac{p_{CN}}{p_{NN}}\Delta^* = \frac{p_{CC}}{p_{NN}}[\overline{Y}(Y \geq y_{1-\alpha_{CC}}) - \overline{Y}^{000}]$$

THen

$$\varepsilon - \varepsilon^* = \overline{Y}^{100} - \overline{Y}(y^{000}_{\alpha_{CN}} \leq Y \leq y^{000}_{\alpha_{CN}+\alpha_{NN}}) - \frac{p_{CC}}{p_{NN}}[\overline{Y}(Y \geq y_{1-\alpha_{CC}}) - \overline{Y}^{000}]$$

Constraint for the late assumptions :

$$\overline{Y}(Y \leq y_{\alpha_{NN}}) < \overline{Y}^{100} < \overline{Y}(Y \geq y_{1-\alpha_{NN}})$$