

How Important is Selection? Experimental Vs Non-experimental Measures of the Income Gains from Migration¹

David McKenzie, *Development Research Group, World Bank**
John Gibson, *University of Waikato*
Steven Stillman, *Motu Economic and Public Policy Research*

Abstract

How much do migrants stand to gain in income from moving across borders? Answering this question is complicated by non-random selection of migrants from the general population, which makes it hard to obtain an appropriate comparison group of non-migrants. New Zealand allows a quota of Tongans to immigrate each year with a random ballot used to choose amongst the excess number of applicants. A unique survey conducted by the authors allows experimental estimates of the income gains from migration to be obtained by comparing the incomes of migrants to those who applied to migrate, but whose names were not drawn in the ballot, after allowing for the effect of non-compliance among some of those whose names were drawn. We also conducted a survey of individuals who did not apply for the ballot. Comparing this non-applicant group to the migrants enables assessment of the degree to which non-experimental methods can provide an unbiased estimate of the income gains from migration. We find evidence of migrants being positively selected in terms of both observed and unobserved skills. As a result, non-experimental methods other than instrumental variables are found to overstate the gains from migration by 20 to 82 percent, with difference-in-differences and bias-adjusted matching estimators performing best among the alternatives to instrumental variables.

Keywords: Migration, Selection, Natural Experiment
JEL codes: J61, F22, C21

¹ We thank the Government of the Kingdom of Tonga for permission to conduct the survey and to use the Tongan Labour Force Survey, the New Zealand Department of Labour Immigration Services for providing the sampling frame, Halahingano Rohorua and her assistants for excellent work conducting the survey, and most especially the survey respondents. Mary Adams, Alan de Brauw, Deborah Cobb-Clark, Chirok Han, Manjula Luthria, Martin Ravallion, Ed Vytlačil and participants at various seminars provided helpful comments. Orazio Attanasio and two anonymous referees provided comments which greatly improved the initial version of the paper. Financial support from the World Bank, Stanford University, the Waikato Management School and Marsden Fund grant UOW0503 is gratefully acknowledged. The views expressed here are those of the authors alone and do not necessarily reflect the opinions of the World Bank, the New Zealand Department of Labour, or the Government of Tonga.

* Corresponding author: E-mail: dmckenzie@worldbank.org. Address: MSN MC3-307, The World Bank, 1818 H Street N.W., Washington D.C. 20433, USA. Phone: (202) 458-9332, Fax (202) 522-3518.

1. Introduction.

How much do migrants stand to gain in income from moving across borders? The answer to this question is helpful in addressing many fundamental economic questions, such as the sources of differences in income levels across countries. It is also a crucial input into calculations about the global gains from more migration, which are a backdrop to policy discussions.² In an influential study, Walmsley and Winters (2003) used wage differentials as a measure of the income gain from migration and estimated that a 3% increase in migration from developing countries would lead to a gain in world income greatly exceeding the gains to be had from removing all remaining barriers to goods trade.

However, even estimates of the income gains from migration that go beyond simple cross-country comparisons of wage rates are likely to be misleading. Ideally, one must compare the earnings of the migrant to what they would have earned in their home country, but the latter is unobserved. Simple comparisons of movers and stayers are therefore likely to be misleading, as income gains may just reflect unobserved differences in ability, skills, and motivation, rather than the act of moving itself. While statistical corrections for non-random selection are often used when modelling migration (Borjas, Bronars and Trejo, 1992), there are doubts about the assumptions behind these methods (Deaton, 1997). These doubts persist because it is hard to know how well these remedies compare with the ideal of a randomized experiment.

This paper uses a unique random selection mechanism to overcome the interpretation difficulties posed by the non-random selection of migrants. The Pacific

² Of course, these policy discussions are also heavily influenced by the very large literature on the impacts of immigration on natives.

Access Category (PAC) under New Zealand's immigration policy allows an annual quota of Tongans to migrate permanently to New Zealand. Many more applications are received than the quota allows, so a random ballot is used by the New Zealand Department of Labour to select from amongst individuals who register. A survey administered by the authors collects data on winners and losers in this ballot. Thus, we have a group of migrants and a comparison group who are similar to the migrants, but remain in Tonga only because they were not successful in the ballot.

By comparing outcomes for these groups, we are able to obtain the only known experimental measure of the average gain in income from migration. As not all of the individuals whose names were selected in the ballot had migrated by the time of our survey, this estimate accounts for non-compliance to the "treatment" of migration. We estimate that there is a 263% increase in income from migrating, measured after only one year in New Zealand for the migrants. While this gain in income is large, it is only half of what a simple comparison of differences in per capita GDP would predict and only 43% of the difference in manufacturing wages between the two countries.

In addition to winners and losers in the PAC ballot, we also surveyed non-applicants. We use this sample along with the migrant sample to obtain non-experimental estimates of the income gains from migration. Five popular non-experimental methods for dealing with selectivity are considered: first-differences, OLS, difference-in-differences, matching and instrumental variables. Instrumental variables using pre-migration distance to the immigration office in Tonga performs best, coming within 1% of the experimental estimate. The other non-experimental methods are found to overstate the gain in income from migration compared to the experimental estimate. OLS, first-

differences and matching without controlling for past income all overstate the income gains by 25-35%, and are statistically different from the experimental estimate. Apart from the distance instrumental variable estimator, we find that a flexible difference-in-differences specification and the bias-adjusted matching estimator of Abadie and Imbens (2006) using past income perform next best, but they both overstate the gain in income by about 20%, which is marginally significantly different from the actual gain.

This paper contributes to two important literatures. First, the existing empirical literature on migrant selectivity focuses on observable measures of skills, such as education (e.g. Chiquiar and Hanson, 2005). While we do indeed see positive selection of Tongan migrants in terms of observed skills, the overstatement of the income gains by the non-experimental methods indicates that Tongan migrants are also positively selected in terms of unobserved ability. Thus, we extend the existing literature by using pre-migration earnings to look at selection on what is typically an unobservable. We find that Tongan migrants are also positively selected in terms of this unobserved component of labor earnings, even after controlling for age, education and other observed characteristics of individuals.

Second, since the influential work of Lalonde (1986), a literature attempts to assess the ability of non-experimental estimators to obtain estimates similar to experimental results. After Lalonde's initial pessimistic assessment of non-experimental measures, there has been much recent debate as to the ability of propensity-score matching methods to obtain better results (e.g. Heckman, Ichimura and Todd, 1997; Dehejia and Wahba 2002; Smith and Todd 2005a; Dehejia 2005). To date, this literature

has concentrated on a small number of labor market training programs.³ Our paper extends this literature in a number of ways. First, the size of the “treatment” considered here is strongly significant and is at least an order of magnitude larger than in Lalonde’s study (and in the majority of papers on other training programs). Second, we explicitly test whether the non-experimental estimates are statistically different from those of the experiment, something that is not done in the preceding studies. Third, we compare the relative performance of different non-experimental estimators in terms of their bias, in order to obtain recommendations on which perform best. Even though our example has many of the features identified by past studies as conducive to more accurate non-experimental estimation (such as a common survey design and a measure of pre-treatment outcomes), we find that non-experimental estimators still overstate the income gains from migration.

The rest of this paper is structured as follows. Section 2 describes the immigration process used as the natural experiment and the sampling method and data from the Pacific Island-New Zealand Migration Study (PINZMS). Section 3 constructs the experimental estimates. Section 4 looks directly at selection into migration, Section 5 estimates five different types of non-experimental estimates and Section 6 concludes.

2. The Pacific Access Category and PINZMS Data

2.1. The Pacific Access Category

The natural experiment we use is based on the Pacific Access Category (PAC) under New Zealand’s immigration policy. The PAC was established in 2001 and allows an annual quota of 250 Tongans to migrate as permanent residents to New Zealand

³ Glewwe, Kremer, Moulin and Zitzewitz (2004) is an exception, comparing regression and difference-in-difference estimates to the results of a randomized experiment on the effects of providing flip charts in schools in Kenya. However, they do not consider matching or IV methods as alternatives.

without going through the usual migration categories used for groups such as skilled migrants and business investors.⁴ During the period we consider, the process of migrating through the PAC was as follows.

Interested individuals must first lodge an application form at the New Zealand Department of Labour (DoL) office in Nuku'alofa, the capital city of Tonga during the one month per year registration period. To be eligible to register, they must be Tongan citizens aged between 18 and 45, and meet certain English, health and character requirements.⁵ The person who registers is the Principal Applicant. If they are successful, their spouse and dependent children are eligible to migrate as Secondary Applicants. The quota of 250 applies to the total of Primary and Secondary Applicants, and corresponds to about 70 migrant households per year.

Since many more applications are received than the quota allows, an electronic ballot is used by the DoL to randomly select from amongst the registrations. Successful ballots are notified by the DoL and then have six months to fill out an application for permanent residency. At this step, applicants are required to provide evidence of a valid job offer in New Zealand. The job offer is required since migrants are meant to be self-sufficient in New Zealand, given that Tongans need two years of residency in New Zealand before becoming eligible for most welfare benefits. The typical jobs taken by the migrants in our sample are low-skilled entry jobs typical of many developing country

⁴ The Pacific Access Category also provides quotas for 75 citizens from Kiribati, 75 citizens from Tuvalu, and, prior to the December 2006 coup, 250 citizens from Fiji to migrate to New Zealand. There have been some changes in the conditions for migration under the Pacific Access Category since the period we examine in this paper (see Gibson and McKenzie (2007) for details) – here we describe the conditions that applied for the potential migrants studied in this paper.

⁵ These include having no criminal record, not having been deported from New Zealand, and not having a highly infectious disease. Data supplied by the DoL for residence decisions between November 2002 and October 2004 show only 1 person was rejected for failing the English requirement and 3 others for failing other requirements of the policy.

migrants, such as sales assistants and packers in retail sales, and carpenters, technicians and welders in building and construction.

After a job offer is filed along with their residence application, it typically takes three to nine months for an applicant to receive a decision. Once receiving approval, they are then given up to one year to move. The median migrant in our sample moved within one month of receiving their residence approval.

As discussed in more detail below, our survey samples from the first three years of the PAC, consisting of the 2002, 2003 and 2004 ballots. Combining these three years, there were 278 successful ballots out of 2,632 registrations, giving a success rate of 10.6%. These 2,632 registrations represented 2,194 individuals, since some registered in multiple years. Based on the 1996 Tongan Census, 36,500 18 to 45 year olds are estimated to be living in Tonga in 2004.⁶ Thus, approximately 6% of the age-eligible Tongan population applied to the PAC during this period.

The other options available for Tongans to migrate are fairly limited, unless they have close family members abroad. Almost all (94%) Tongan migrants live in New Zealand, the United States and Australia.⁷ In the 2004/05 financial year, New Zealand admitted 58 Tongans through a business/skilled category and 549 through family sponsored categories.⁸ Australia admitted 284 Tongans during the same financial year.⁹ The United States admitted 324 Tongans in the 2004 calendar year, comprising of only five under employment-based preferences and 290 under immediate relative or family-

⁶ Source: Tonga Statistics Department, Key Statistics Excel Worksheet.

⁷ Source: GTAP database of Parsons et al. (2005).

⁸ Source: Residence Decisions by Financial Year datasheet provided by DoL. Migrants under the family sponsored categories were mainly parents and spouses.

⁹ Source: Settler Arrivals 2004-2005, Australian Government Department of Immigration and Multicultural Affairs.

sponsored categories.¹⁰ The main alternative to the PAC is thus family categories of migration, which in practice are mostly limited to spouses and parents of existing migrants.

2.2. The Pacific Island-New Zealand Migration Survey

The Tongan component of the Pacific Island-New Zealand Migration Survey (PINZMS), is a comprehensive household survey designed to take advantage of the natural experiment provided by the PAC.¹¹ The survey design and enumeration, which was overseen by the authors in 2005, covered random samples of four groups, surveying in both New Zealand and Tonga.

The first group consists of Tongan immigrants in New Zealand, who were successful participants in the 2002-2004 PAC ballots. The sample frame for this group was the names and addresses of the 92 successful ballots who were approved for residence and had arrived in New Zealand as of January 8, 2005.¹² We conducted our fieldwork over several months in Auckland and surrounding areas (including Hamilton), where most Pacific Island migrants locate, and made single visits to the second and third largest cities of Wellington and Christchurch. This fieldwork located 66 of the 92 migrants, only one of whom refused to be surveyed. Of the remaining 26, 6 were in areas outside our survey area, 11 were untraceable because the only address details they gave on their application was a P.O. Box in Tonga for the immigration agent they used.¹³ The

¹⁰ Source: 2004 Yearbook of Immigration Statistics, US Department of Homeland Security Office of Immigration Statistics.

¹¹ See www.pacificmigration.ac.nz for additional details and further papers based on this survey.

¹² This information was supplied under a contractual arrangement with the DoL, who can use passport records to detect which dates individuals enter and leave New Zealand. None of the migrants had returned to Tonga at the time of the survey, nor had they as of May 2007, two years after the survey.

¹³ For the migrants who were traced, the most useful tracing information came from their remaining family and friends in Tonga, or from others in their village. Such contacts were unknown for those who only supplied an agent P.O. Box rather than an address containing at least the village of residence.

remaining 9 are likely to be cases where the formal names on passport and PAC records differ from the names by which people are known amongst their community. Thus even though we used Tongan interviewers with extensive community contacts, it was not possible to find these people. The data on the application forms is very limited, preventing a detailed analysis of those who we could not locate, but we are able to check and confirm that they do not differ significantly from the migrants in our sample in terms of the proportion who are male, the date at which their residence decision was approved and their last date of entry in New Zealand. We will nonetheless examine the robustness of the estimated income gain from migration to alternative assumptions about the characteristics of these non-surveyed individuals.

The second group consists of successful participants from the same random ballots who were still in Tonga at the time of surveying. The sampling frame for this group was the names, addresses and telephone numbers of the 186 successful ballots who were still living in Tonga at the time of the survey. Budget constraints led us to draw a random sample of 55 from this group, located in villages from which the migrants in our first survey group had emigrated. This group includes individuals whose applications were still being processed at the time of surveying, a few whose applications for residence were rejected for lack of a valid job offer, and individuals who did not end up applying to emigrate after their ballots were chosen. Most are in the first category. Our latest update from the DoL shows that, as of May 2007, 235 out of the 278 successful ballots (85%) are now in New Zealand. In forming our experimental estimate, we weight the sample so that it reflects the actual ratio of migrants to successful ballots still in Tonga.

The third survey group consists of unsuccessful participants in these same ballots. The full list of unsuccessful ballots from these years was provided to us by the DoL, but the details for this group were less informative than those for the successful ballots, as only a post office box address was supplied and there were no telephone numbers. We used two strategies to derive a sample of 78 unsuccessful ballots from this list, with this sample size again dictated by available budget. First, we used information on the villages where migrants had come from to draw a sample of unsuccessful ballots from the same villages (implicitly using the village of residence as a stratifying variable). Second, we used the Tongan telephone directory to find contact details for people on the list. To overcome concerns that this would bias the sample to the main island of Tongatapu, where people are more likely to have telephones, we deliberately included in the sample households from the Outer Islands of Vava'u and 'Eua.

The final survey group consists of households living in the same villages as the PAC applicants but from which no eligible individuals applied for the quota in any of our sample years (e.g. 2002-2004). We randomly selected 60 non-applicant households with at least one member aged 18 to 45, giving an overall sample of 183 18 to 45 year olds. This group is used for our non-experimental comparisons. For all three survey groups in Tonga, we had no cases of people refusing to take part in the survey.

The same survey instrument was then administered to all four groups by the same survey team. The survey collected data on employment, income and demographics for all household members, along with detailed modules on health, remittances and the migration experience. Income is measured as the before-tax weekly income from wages, salaries, commissions and non-agricultural businesses. Tongan pa'anga were converted

into New Zealand dollars at the prevailing market exchange rate of 1 Tongan pa'anga to 0.729 New Zealand dollars.¹⁴ Our own calculations suggest that the PPP exchange rate at the time was very close to this (see McKenzie, Gibson and Stillman, 2006).

At the time of the survey, the median migrant had spent 14 months in New Zealand (mean 12 months), with the median migrant arriving in 2004 (and the earliest arriving in 2003). Their earnings are therefore informative as to what individuals can earn merely by crossing a border, before they have had time to learn local skills and assimilate. Migrants were also asked what their main occupation was in the 12 months prior to moving to New Zealand and their usual weekly income in Tonga in the first half of the year that they migrated (e.g. their past income). For the median migrant, past income therefore refers to their usual work income at the start of 2004. Individuals surveyed in Tonga were asked retrospective questions on their usual work income in the first half of 2004 and first half of 2003. We use the 2004 report as our measure of past income to correspond to the median immigrant. However, our results are robust to using the average of 2003 and 2004 income as the measure of past income for non-migrants.

2.3. Verifying Randomization

Table 1 examines the quality of our samples (i.e. whether they are truly random) by comparing the means of ex-ante characteristics for ballot winners and losers among the principal applicants in our sample. The point estimates are similar in magnitude for the two groups and we can not reject equality of means for any of the variables. In particular, the two groups have similar years of education, age, marital status and birthplace. Principal applicants in both groups were each, on average, earning \$NZ77-81 per week in Tonga in early 2004. These results indicate that there is, in fact, random

¹⁴ At the time of the survey NZ\$1 = US\$0.72.

selection of ballots among applicants to the Pacific Access Category and that our surveying strategy has generated random samples of ballot winners and losers, and thus the sample of unsuccessful ballots are a proper comparison group for estimating the impact of migration on the income of successful principal applicants.

3. Experimental Estimates of the Income Gains from Migration

To determine the income gains from migration, one must compare the earnings of each migrant to what they would have earned in their home country had they not migrated. Typically, it is not possible to identify this unobserved counterfactual outcome. However, the PAC ballot system, by randomly denying eager applicants the right to move to New Zealand, creates a control group of individuals that should have the same outcomes as what the migrants would have had if they had not moved. The mean weekly income of PAC migrants is \$NZ424 in our sample, compared to \$NZ104 for unsuccessful ballots. This simple comparison of means therefore suggests that Tongans gain \$NZ320 per week in labour earnings from migrating to New Zealand.

However, as discussed in Heckman et al. (2000), a simple comparison of the treatment and control groups in an experimental context will be biased if control group members substitute for the treatment with a similar program or if treatment group members drop out of the experiment. In our application, *substitution* bias will occur if PAC applicants with unsuccessful ballots migrate to New Zealand through an alternative visa category such as the family or skills category or migrate to another country and *dropout* bias will occur if PAC applicants whose names are drawn in the ballot fail to migrate to New Zealand. We do not believe that substitution bias is of serious concern in

our study, as individuals with the ability to migrate via other arrangements will likely have done so previously given the low odds of winning the PAC ballot.¹⁵

However, dropout bias is a more relevant concern, since only one-third of ballot winning principal applicants had migrated to New Zealand at the time of our survey. The estimated income gain of \$NZ320 will only be a consistent estimate of the gains from migration if there is no selection as to who migrates among those successful in the ballot or if the return to migration is homogeneous. If the ballot winners who will earn relatively higher incomes after migration are more likely to migrate, then comparing migrants to ballot losers will overstate the income gains from migration. We therefore turn our attention to measures of the impact of migration which are consistent even if there is selective migration among those with successful ballots.

3.1. Experimental Estimates

We can use our survey data to recover two parameters of interest. The first is the mean impact on income of winning the PAC ballot, known as the intention-to-treat effect (ITT). This estimator can be computed by comparing the mean income for ballot winners to that for ballot losers, which shows an increase in weekly income of \$NZ91. To increase the precision of our estimate, we re-estimate the ITT using the OLS regression model described in equation (1) and add controls for the observable pre-existing characteristics of the two groups:

$$\text{Income}_i = \alpha + \beta * \text{BallotSuccess}_i + \delta' X_i + \omega_i \quad (1)$$

Column 1 of Table 2 reports the results from estimating this regression with no controls, repeating the estimate of \$91 obtained as the difference in means. In Column 2,

¹⁵ We did not come across any incidences during our fieldwork where remaining family members told us that the unsuccessful applicant had migrated overseas.

we add as controls standard wage equation variables, such as age, sex, marital status and years of education. In addition, we include height, as a pre-existing measure of health, and whether or not the applicant was born on the main island of Tongatapu, as a measure of having more urban skills. Adding these controls reduces the size of the estimated effect only slightly, to \$90, which is not significantly different from the original estimate. Column 3 adds further controls for past income, which is expected to capture the effect of a host of unobserved individual attributes that determine income. This only marginally changes the estimated intent-to-treat effect, which is now estimated to be \$87. The fact that the estimated program effect changes only slightly in magnitude as we add controls is unsurprising given that the results in Table 1 show that the PAC ballot and our survey provides a proper randomized sample.

The ITT measures the impact of receiving a successful PAC ballot, rather than the impact of migration, which is our main object of interest. However, we can estimate the impact of migration by using the outcome of the PAC ballot as an instrument for migration when estimating equation (1) modified to have migration rather than ballot success as the regressor of interest. This provides the local average treatment effect (IV-LATE), which is interpreted as the effect of treatment on individuals whose treatment status is changed by the instrument. In our application, this is the effect of migration on the income of individuals who migrate after winning the ballot. Angrist (2004) demonstrates that in situations, such as ours, where no individuals who are assigned to the control group receive the treatment (e.g. there is no substitution), the IV-LATE is the same as the average treatment effect on the treated (IV-TT).

Having a successful ballot is of course strongly correlated with migration (the first stage F-statistic is 61.5). Validity of the exclusion restrictions then requires: (i) that success in the ballot is uncorrelated with individual attributes which might also affect income, which is provided by the randomization of the ballot draws; and (ii) that the ballot outcome does not directly affect incomes, conditional on migration status. One could conceive of stories such as that winning the ballot and not being able to migrate causes frustration and leads individuals to work less, or conversely, that winning the ballot acts as a spur to work harder in order to afford the costs of trying to find a job in New Zealand. However, we did not encounter any evidence of such changes in behaviour in our field work, lending support to this identification assumption.

Column 4 of Table 2 reports the IV-TT estimator when no other controls are included in the regression model. We estimate a gain in weekly work income of almost \$274 from migrating. Column 5 adds the same control variables used above when estimating the ITT, with the estimated gain increasing slightly to \$281. Column 6 then adds past income as a further control, which results in an estimated income gain from migration of \$274 per week.

Therefore, controlling for the pre-existing characteristics of the treatment and control groups, we estimate that a successful ballot increases the mean income of PAC applicants by \$87 per week, while migrating increases mean income by \$274. Given that the mean income of applicants with unsuccessful ballots is \$104, this represents a 84% increase in income from winning the ballot and a 263% increase in income from migrating. While large, this gain is much less than that predicted by differences in per capita income. In 2004, New Zealand's GDP per capita was NZ\$30,469, while Tonga's

was NZ\$2,044.¹⁶ This difference equates to NZ\$546 per week which is twice as large as the actual gain experienced by the average migrant in our survey. The difference in manufacturing wages of NZ\$635 is even larger, with the experimental estimate of the income gain only 43% of this difference.¹⁷

3.2 Sensitivity Analysis to Unsurveyed Migrants

As noted above, we were able to locate approximately 70 percent of the migrants in New Zealand. The reasons why individuals were not located do not suggest that these individuals are particularly different from those in our survey, and they are indeed similar in terms of the few characteristics included on their application forms. Moreover, the randomization tests in Table 1 suggest that the *sample* of PAC ballot winners is indeed comparable to that of PAC ballot losers in terms of many more characteristics, and our experimental estimates do not change when such controls are added. For these reasons, we believe the experimental estimate of the income gain from migration is not biased by the exclusion of migrants that we could not locate.

Nevertheless, we can examine the sensitivity of our estimated experimental income gain from migration to different extreme assumptions about the income that these non-surveyed migrants earn in New Zealand. As an extreme lower bound, we assume that only 70 percent of these non-surveyed migrants work, and that those who work each earn NZ\$360 per week, the 10th percentile of the weekly income distribution among principal applicant migrants in our sample. With these assumptions, the estimated TOT is \$221. As

¹⁶ Source: World Bank GDF and WDI Central (August 2005 update) for population and GDP.

¹⁷ This calculation uses average manufacturing weekly income from the Tongan Manufacturing Census in 2002 (www.spc.int/prism/Country/To/stats/Economic/Production/Manufacturing/wages_salaries.htm) and the New Zealand Quarterly Employment Survey (averaged over 2002), converts Tongan pa'anga to New Zealand dollars at the 2002 average exchange rate, and then uses the New Zealand consumer price index to convert 2002 dollars to March 2005 dollars.

an extreme upper bound, we assume that 100 percent of the non-surveyed migrants are employed, and that they each earn \$748 a week, the 90th percentile of the earnings distribution for migrants in our sample. In this case the estimated TOT is \$368.

These extreme assumptions then give bounds on the income gain from migration of 213% to 354%. So, while the available evidence suggests that the non-surveyed migrants are indeed similar to those surveyed, even under extreme assumptions, income gains from migration are large. Moreover, since we are using the same migrant sample when applying each of the non-experimental estimators, our comparison of the difference between experimental and non-experimental estimates will not be affected by any non-random sampling from the group of successful ballots.¹⁸

4. The Self-Selection of Migrants

An extensive literature examines migrant selectivity. Most datasets on migrants lack information on their earnings prior to migration, leading much of this literature to focus on comparing observable characteristics of migrants to those of non-migrants (e.g. Borjas 1987, Chiquiar and Hanson 2005). The average PAC ballot entrant in our sample has 11.9 years of education, compared to 9.8 years among 18 to 45 year-old non-applicants, showing positive selection of Tongan migrants in terms of observable skill. However, the main concern with using non-experimental estimators to measure the income gains from migration is that migrants also differ from non-migrants in terms of unobserved qualities. Using our data, we can examine differences in pre-migration earnings to establish whether there is positive or negative selection on unobservables

¹⁸ That is, since we are using the same migrant sample for both experimental and non-experimental estimators, if we understate the mean income of migrants due to survey non-response, this will lower the income gain estimated from the experiment, and also the income gain estimated from non-experimental methods by the same amount.

conditional on the characteristics typically found in survey data. We can also examine whether unobservables are related to selective compliance among the PAC ballot winners.

We first examine the overall extent of selection by comparing the pre-migration earnings of migrants to that of observationally similar non-applicants via the following regression:

$$\text{Income}_{i,t-1} = \alpha + \beta * \text{Migrant}_{i,t} + \gamma' X_{i,t} + \varepsilon_{i,t-1} \quad (2)$$

where X consists of a set of time-invariant controls, such as age, education, gender, marital status, height, and migrant network, and $Migrant$ is a dummy variable taking the value one if person i applies for the ballot and migrates and zero if they don't apply for the ballot. The coefficient β then indicates whether migrants earned more or less prior to migration than non-applicants, conditional on their observed characteristics.

The first two columns of Table 3 report the results of estimating equation (2), comparing migrants to all 18-45 year-old non-applicants. The coefficient β is positive and highly significant. Migrants and non-applicants are seen to differ both in terms of observables and unobservables. Without any controls, migrants are found to earn \$56 per week more than non-applicants prior to emigrating. Controlling for observables lowers this difference to \$31 per week. Given that the average income of non-applicants is \$34 per week, migrants are estimated to have earned almost twice as much as observationally similar non-applicants prior to emigrating. This indicates that there is positive selection on unobservables which should lead non-experimental estimators to overstate the income gains from migration; something that we will test in the next section.

There is a smaller effect of unobservables on selective compliance among those who win the PAC ballot. Modifying equation (1) to compare the pre-migration incomes of ballot winners who migrate to ballot winners who had not migrated at the time of the survey gives statistically insignificant coefficients of \$13 per week without controls and \$5 a week with controls. While there is less evidence of selection here, the point estimates are consistent with the NZ\$320 comparison of migrants to ballot losers being an overestimate of the experimental IV-TT of NZ\$274.

5. A Comparison of Non-Experimental Estimators

The natural experiment provided by the PAC quota provides a unique opportunity to estimate the gain in income from migration. Other studies of migration are forced to use non-experimental methods to attempt to deal with the selectivity issues associated with migration, comparing the incomes of migrants to that of non-migrants of similar observable characteristics. These methods will only produce unbiased estimates of the impact of migration if all differences between migrants and non-migrants are captured in survey questions or there is an instrumental variable that is highly correlated with migration status, but does not affect incomes conditional on migration status. In this section, we explore how well such methods work in practice, comparing the results obtained from different non-experimental methods to the experimental results described above.

5.1. Related Literature

This approach to studying the validity of non-experimental methods has a long history in the program evaluation literature. For example, Lalonde (1986) compared experimental estimates from the National Supported Work (NSW) Demonstration to non-

experimental results calculated using control groups created from household survey data. His male experimental group consists of 297 treated individuals and 425 controls, and the experimental estimate of the change in earnings ranged from \$851-883 with a standard error of 306-323, relative to baseline earnings of \$3,000. While this treatment effect of 28% was statistically significant, after controlling for observables (his Table 5, column 10), the treatment effect falls to \$662 with a standard error of \$506 – i.e. it is insignificantly different from zero.

For this program and treatment, Lalonde found that non-experimental methods did a poor job of replicating the experimental results. As comparison groups he used the large multi-purpose PSID and CPS surveys. The non-experimental control group sample sizes ranged from 15,992 when all individuals in these broad surveys were included, down to 128 in his PSID-3 sample, when he attempted to chose men more similar to those in the experiment. The estimated treatment effect, after controlling for all observables and for income, ranged from –1,228 to +1,466 depending on the control group, which he took as evidence that the non-experimental estimators did not give accurate or precise estimates. However, he did not test whether the non-experimental estimates were statistically different from the experimental estimate.

Dehejia and Wahba (2002) and Smith and Todd (2005a,b) use the same NSW data as Lalonde (1986) to examine whether propensity-score matching techniques yield results which are closer to the experimental estimates. Dehejia and Wahba (2002) further restrict the experimental sample to 185 treated individuals and 260 control individuals in order to obtain two years of pre-treatment earnings information. Their matched sample from the PSID then ranges from 56 to 2,021 observations depending on the caliper used.

Using nearest neighbor matching, with replacement, they find treatment effects of \$1,824-\$1,973, compared to their experimental estimate of \$1,794, which they view as evidence that non-experimental methods can produce accurate estimates, although they do not explicitly test whether there is a significant difference between the experimental and matching estimates. However, Smith and Todd (2005a,b) show that their results are very sensitive to the choice of sample, propensity score specification and tie-breaking rule used for choosing the closest match.

This literature has concluded that more accurate non-experimental estimates can be achieved if the treatment and non-experimental control groups are: i) compared over a common support (e.g. the distribution of the likelihood of receiving the treatment is similar in both groups), ii) located in the same labour markets, and iii) administered the same questionnaire (e.g. data is collected from both groups in an identical manner). Further improvement can be achieved if data are collected from both the pre- and post-treatment periods and a ‘difference-in-differences’ estimator is used to control for unobserved differences between the treatment and control groups by differencing out individual fixed effects which are correlated with both the outcome and the likelihood of being treated. Nonetheless, even with these refinements, Smith and Todd (2005, p.305) conclude, “Our analysis demonstrates that while propensity score matching is a potentially useful econometric tool, it does not represent a general solution to the evaluation problem.”

5.2. Non-Experimental Control Groups with Tongan Data

The PINZMS collected data for a sample of non-applicants to the PAC ballot selected from either the same villages that the migrants had been living in prior to

migrating or in the same villages that unsuccessful ballots were found in and used the same survey team to administer an identical questionnaire to the one given to other non-migrants in our sample (eg. the experimental control group). Thus, these individuals serve as an ideal non-experimental control group on which to test alternative methodologies for estimating the gains from migration. As discussed above, all individuals in our sample report their income from the previous year allowing us to also implement a ‘difference-in-differences’ estimator.

The PINZMS non-applicant sample of 180 individuals aged 18 to 45 is similar in size to some of the samples used by Lalonde (1986) and Dehejia and Wahba (2002). For robustness, we also use a large general survey to form the control group for our non-experimental estimates, as these previous studies have done with the PSID. The 2003 Tongan Labor Force Survey (TLFS) is the most recent Labor Force Survey in Tonga.¹⁹ It is a nationally representative survey of 8,299 individuals, of whom 3,979 are aged 18-45 and contains information on income, age, sex, marital status, education and birthplace. However, the survey does not measure height, collect data on past income, or the variables that we use for instruments. Thus, we use the TLFS only to test the robustness of the OLS and propensity-score matching estimators.

5.3. Testing for Significant Differences between the Non-Experimental and Experimental Estimates

We use the non-applicant sample from the PINZMS and the TLFS along with the migrant group in the PINZMS to calculate different non-experimental estimates of the

¹⁹ See Tonga Statistics Office (2004) for a summary report and detailed description of the survey methodology. To use this data we convert 2003 Tongan Pa’anga into 2005 New Zealand dollars using the Tongan Consumer Price Index to convert 2003 Pa’anga into 2005 Pa’anga and the first quarter 2005 exchange rate to convert to New Zealand dollars.

income gain from migration. These non-experimental estimates are then compared to the experimental estimate, obtained from the PINZMS sample of ballot applicants. An issue which arises when comparing these estimates is that they are not estimated from independent samples. Instead, the same sample of migrants is used to construct all of the estimates, while the same PINZMS sample of non-applicants is used for the majority of non-experimental estimates. This is an issue that has been ignored in the previous literature (Lalonde, 1986; Dehejia and Wahba, 2002; Smith and Todd, 2005a,b).

We develop a bootstrap procedure that accounts for the fact that multiple estimates are being constructed from the same sample and allows us to formally test whether the non-experimental estimates are different from the experimental estimate. The procedure works as follows: First, we divide the data into five groups: (a) Migrants, (b) Non-compliers in the PINZMS data, (c) Unsuccessful ballots in the PINZMS data; (d) Non-applicants in the PINZMS data; and (e) 18-45 year-olds in the TLFS. Then, we sample with replacement from each of these groups and pool these new samples together. Next, the entire set of experimental and non-experimental estimators are estimated using this new sample and the results are saved. Finally, we replicate this procedure 1000 times and calculate the percentage of bootstrap replications in which a particular non-experimental estimate exceeds the experimental estimate. This provides a one-sided test of the difference in the estimated treatment effect.²⁰ For comparison, we also calculate a 90% confidence interval for the difference between each non-experimental estimate and the experimental estimate, which is also used to compare the relative performance of the different non-experimental estimators.

²⁰ This one-sided test is appropriate since the results in section 4 indicate that migrants are positively selected in terms of unobservable characteristics in which case non-experimental methods will tend to overstate the gains.

5.4. The Single Difference Estimator

We begin by examining whether a simple single difference estimate calculated using only information from the migrant group provides a good estimate of the income gains from migration. Several recent surveys of new immigrants (e.g. the Longitudinal Immigrant Survey: New Zealand and the US New Immigrant Survey) ask about income prior to migration. Thus, one approach to estimating the average income gain from migration is to calculate the mean difference between the migrant's pre-migration and post-migration incomes.

There are several possible sources of bias in such an estimate. The proper counterfactual is what a given individual would be earning in the current time period if he or she did not migrate; this could be different from what they earned before migration due to macroeconomic factors, such as aggregate growth, or because of changes in the earning potential of the individual over time, perhaps because of returns to labor market experience. An additional potential form of bias is that the recall of previous income may involve omissions or telescoping errors, leading to non-mean zero measurement error.

The first row of Table 4 reports the single-difference estimate, calculated as the difference between the current income of our migrant sample and what they reported earning prior to migration. Using this method, we estimate an income gain of \$341. Comparing this to the experimental estimate of \$274 (column 4, Table 2), the estimated income gain is overstated by 25%. The single-difference estimate is greater than the experimental estimate in 92.7% of our replications. A one-sided t-test therefore rejects equality of the two estimates at the 10% level.

5.5. OLS

A second non-experimental method commonly used to estimate the returns from migration is to assume that all differences between migrants and non-migrants that affect income can be captured by the control variables in an OLS regression. One then estimates λ through the following regression:

$$\text{Income}_i = \kappa + \lambda * \text{Migrate}_i + \pi' X_i + \upsilon_i \quad (3)$$

We estimate equation (3) by combining the sample of migrants in New Zealand with either the PINZMS non-applicants or the TLFS 18 to 45 year-olds. We consider three sets of control variables. Using the PINZMS non-applicants, the basic specification includes the same controls as those used for the experimental estimates, with the exception of past income (which we keep for the difference-in-differences estimator below). We then allow for a more flexible specification by interacting the male dummy variable with each of the other regressors in the base specification and including fourth order polynomials in age and years of education, along with the interaction of age and years of education. An F-test of joint significance of these additional 12 regressors has a p-value of 0.056. The TLFS data does not include height, so to compare the PINZMS and TLFS results for the same control variables, we repeat these two specifications for the PINZMS non-applicants excluding height and then estimate the same models for the TLFS sample. Appendix 1 provides the full regression results for the base specifications.

Table 4 shows the estimated income gains using OLS. The PINZMS and TLFS control samples give very similar results and including height in the PINZMS does not make much difference. The linear specifications estimate the income gain at \$360-369, which is 31-35% higher than the experimental estimates. Adding polynomial terms

reduces the estimates only slightly, with the point estimates in the \$347-358 range, 27-31% higher than the experimental estimate. The bootstrap replications show that the OLS estimates exceed the experimental estimate in 95-99% of cases. None of the 90% confidence intervals for the difference between the OLS and experimental estimate contain zero. Thus, we can reject equality of the OLS and experimental estimates, and see that OLS overstates the income gains from migration.²¹

5.6. Difference-in-Differences

Using retrospective-reported past income, we can also control for time invariant individual attributes that affect labor market income by estimating a difference-in-differences regression. Since we do not have panel data on all of the control variables, which are likely to be invariant over the short time-period studied, we estimate the following version of the difference-in-differences regression:

$$\text{Income}_i - \text{PastIncome}_i = \kappa + \lambda * \text{Migrate}_i + \pi' X_i + \upsilon_i \quad (4)$$

We again estimate this using both linear and polynomial terms in the controls. Table 4 shows that controlling for past income lowers the estimated income gain to \$330-334, which is still 21-22% higher than the experimental estimate. The estimate from the linear specification is higher than the experimental estimate in 90.8% of bootstrap replications, while the estimate from the polynomial specification exceeds the experiment estimate in 86.9% of cases. The polynomial point estimate is larger but has a higher standard error and thus the difference is marginally significant. The bootstrapping shows

²¹ To further investigate the direction of selection on observables, columns 1 and 2 of Appendix 1 compare the regression results with and without controlling for observable characteristics. Adding basic human capital controls lowers the estimated income gain from \$386 to \$360, which is consistent with positive selection on observables. However, the change in the migration coefficient from adding these controls is not significant, and their addition only reduces the overestimation of the income gains from 41% to 31%

that the difference-in-difference estimator has lower bias than OLS in 96% of replications for the baseline models and 77% of replications for the polynomial models. Adding past income thus controls for some of the source of bias in the OLS estimates.

There are two main possible sources of remaining bias. First, unobserved characteristics like drive and ability may be rewarded differently in the New Zealand and Tongan labor markets so that the individual effects are time-varying. Second, we may be comparing migrants to not-very-similar non-migrants and thus the controls in equation (4) are insufficient to account for differences in the underlying trends in labor income for each group. Matching estimates allow us to do this comparison semi-parametrically and thus minimise the second possible source of remaining bias.

5.7. Matching Estimators

Matching is perhaps the non-experimental evaluation technique which has attracted most research interest in recent years, with proponents claiming that it can replicate experimental benchmarks when used appropriately (Dehejia and Wahba, 2002; Dehejia 2005), although matching only yields a consistent estimate of the treatment effect under the assumption of no selection on unobservables. Ham, Li and Reagan (2006) is a recent application of this technique to (internal) migration. We follow much of the existing literature in focusing our attention on the matching estimate of the average treatment effect for the treated (ATT). This is the estimate that is most directly comparable with the experimental treatment effect.

Matching methods used

The most common form of matching used in the literature is propensity-score matching (Rosenbaum and Rubin, 1983). In this approach a probit equation for the

probability of migrating is estimated and each migrant is then matched to non-applicants with similar predicted probabilities of migration. This enables migrants to be compared to individuals who are similar in terms of observed characteristics, the assumption then being that there is no selection on unobservables. Once the matches are constructed, the gain in income is calculated as the mean income for migrants less the mean income for the matched sample. However, recently Abadie and Imbens (2006) have provided a new bias-adjusted estimator which allows for matching on multiple covariates. This method avoids having to impose parametric assumptions in estimating the propensity score. Moreover, they show their bias-adjusted nearest neighbor estimator to perform well on the same NSW data as used by Dehejia and Wahba (2002).

We focus on nearest neighbor matching, using the STATA software developed by Abadie et al. (2001) to carry out the estimation and bias-adjustment. Single nearest neighbor matching is used for most of our analysis,²² although we also show robustness to using multiple nearest neighbors. For robustness, we also carry out propensity-score matching. The propensity score matching estimators were implemented in STATA using the programs developed by Becker and Ichino (2002) which implement the algorithm suggested in the appendix of Dehejia and Wahba (2002) for carrying out propensity score matching. The programme both estimates a probit equation for the probability of migration, given a set of covariates, and tests that the average propensity scores and the means of the each observable characteristic used in forming the propensity score do not differ between the treated and control units within equally spaced intervals of the

²² If several control observations have the exact same distance to the treated observation, the Abadie et al. (2001) program uses all closest matches in forming the nearest neighbor estimate. Given that Smith and Todd (2005b) question the robustness of propensity-score matching to the choice of tie-breaker rule, this appeals as a choice which does not rely on tie-breaking algorithms.

propensity score.²³ This is a necessary condition for the balancing hypothesis underlying propensity-score matching.²⁴ These balancing tests were satisfied in all our applications.

Is there a pre-ballot earnings dip?

In their studies of labor training programs, Heckman, Ichimura and Todd (1997) and Dehejia and Wahba (2002) note the importance of including information on labor force histories in estimating the probability of participation. A particular concern in labor training programs is the dip in earnings often observed prior to participation in such programs (Ashenfelter, 1978). For this reason, Dehejia (2005) stresses that two or more years of pre-treatment earnings are desirable for use in matching.

The gain in income from migrating is so large relative to incomes in Tonga that it seems less likely that the decision to apply for the PAC ballot will be affected by earnings changes prior to the ballot forms being due. We only have income in the year prior to migration for migrants, but are able to check for a pre-migration ballot dip by comparing unsuccessful ballots to non-applicants. To do this, we match unsuccessful ballots to non-applicants in terms of the same baseline characteristics used in the main matching estimates described below, as well as their real weekly work income in 2002 and 2003. We can then ask whether individuals who would apply for the ballot in early 2005 had lower income in 2004 than similar individuals, with similar incomes in 2002 and 2003, who did not apply for the ballot. The estimated mean difference in weekly income in 2004 from this match is -1.63 pa'anga per week, with a standard error of 11.35. This is

²³ The number of interval is determined by first splitting the sample into 5 equally spaced intervals of the propensity score. If within each interval the average propensity scores of the treated and control observations do differ, then the interval is split in half, and tested again, until in all intervals the average propensity score of the treated and control observations do not differ.

²⁴ If one or more of the characteristics do differ, the balancing properties are not satisfied, and the researcher then enters a less parsimonious specification of the propensity score.

statistically insignificant and amounts to less than 2 percent of mean weekly income for ballot applicants. Therefore it seems we can rule out a pre-ballot dip in earnings, and so controlling for one rather than two years of past income should not greatly affect our results.

Results

We begin with discussion of the Abadie and Imbens (2006) matching results, where matching occurs over multiple covariates. Panel A of Table 5 reports the population ATT and bias-adjusted ATT using the PINZMS when past income is not used in the match. The estimated income gain is \$364 without the bias adjustment and \$350 with it. These estimates are 28-33% higher than the experimental estimate and the bootstrap replications indicate that we can reject equality with the experimental estimate. The bootstrap also shows the value of the bias-adjustment. The bias-adjusted estimate is closer than the unadjusted estimate to the experimental estimate in 93.5% of bootstrap replications.

These base specifications are relatively parsimonious, using only six covariates to form the match. We therefore next include higher order interactions, matching on 17 covariates, and estimate the same specification using both the PINZMS and TLFS data. The point estimates are similar using these two distinct surveys -- \$334 using the PINZMS and \$359 using the TLFS. A one-sided test provides strong evidence that the matching estimate is larger than the experimental estimate in the TLFS sample, with borderline significance in the PINZMS sample.

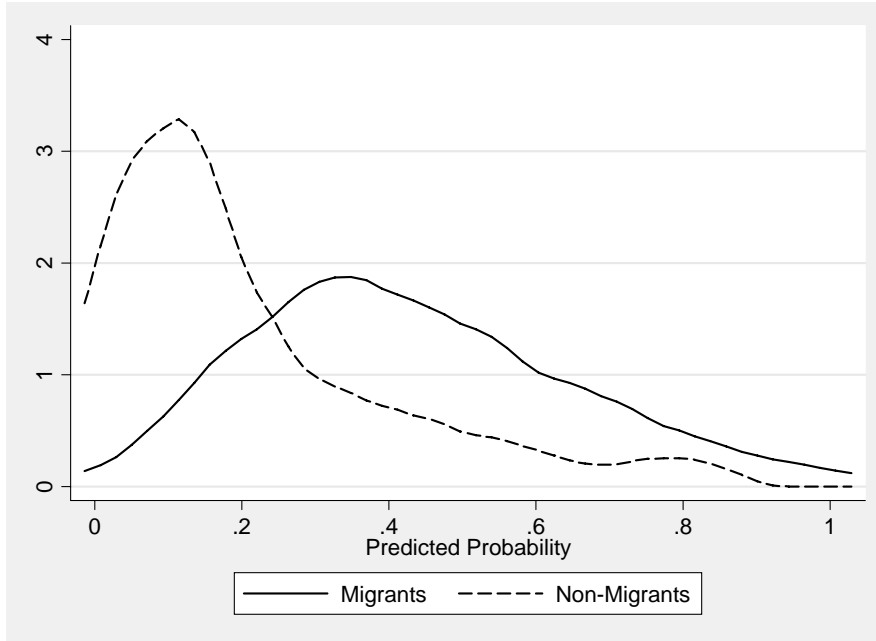
Panels B, C and D of Table 5 explore the sensitivity of the matching estimator to adding controls for past income, adding interactions and to using multiple nearest

neighbors in the matching. Adding past income does move the matching estimate closer to that from the experiment. However, we can still reject equality for the non-bias adjusted estimates. The bias-adjusted estimator with interactions gives an estimated income gain of \$330, which is 20.5% larger than the experimental estimate. This is robust to the number of neighbors used in forming the match. Increasing the number of neighbors used does reduce the standard error, and when multiple nearest neighbors are used, the bias-adjusted ATT exceeds the experimental estimate in over 90% of bootstrap replications.²⁵

As an alternative to matching on multiple covariates, we also consider propensity score matching. We estimate the propensity score as a function of 20 covariates, including past income and an interaction between gender and past income. Figure 1 shows kernel densities of the propensity scores. There is considerable overlap in the distributions, with some migrants and some non-applicants in almost all the range. The propensity score for the migrant group ranges from 0.069 to 0.947, while that of the non-applicant comparison group ranges from 0.000 to 0.789. Estimation is restricted to the area of common support, where the two distributions overlap.

²⁵ We have kept the same set of control variables as used throughout, to enhance comparability across methods. However, matching methods can also condition on potentially endogenous variables. One such variable which is highly correlated with the likelihood of applying to the PAC is the number of relatives an individual has in New Zealand. The bias-adjusted estimate of \$330 actually increases to \$337 when we include this as an additional control indicating that even conditional on network effects, selection on unobservables biases the results.

Figure 1: Propensity Scores for Migrants and Non-migrants



We consider two methods of matching based on the propensity score. The first, radius matching, matches a migrant to all observations whose propensity score does not differ by more than a set radius from the migrant’s propensity score. The second, kernel matching, imposes a Gaussian or Epanechnikov kernel to place more weight on control observations with a closer propensity score. The kernel and radius propensity score matching estimates range from \$330-372, which is 20-36% above the experimental estimate. The kernel matching estimates exceed the experimental estimate in 92% of bootstrap replications.

Finally, we use the TLFS for propensity score matching. The last four rows of Table 5 report results from propensity score matching after trimming observations in the support with very low or very high probabilities of being selected. The same covariates and their interactions are used as for the nearest neighbour matching. We see that trimming the data leads to large numbers of the TLFS control observations being omitted,

since they are very dissimilar to the migrants. After trimming propensity scores below 0.05 or above 0.95, the sample shrinks from 4,043 to 354 observations. In contrast, there is very little change in the sample size when we use the PINZMS non-applicants as controls. This reflects the inherent geographic matching present in the PINZMS sample – controls are chosen from the same villages as migrants, so are more similar to migrants than randomly selected individuals from the nationally representative TLFS. This echoes the conclusions of Smith and Todd (2005b) who find that the vast majority of observations in the CPS and PSID look nothing like the participants in supported work, and so are dropped from their analysis. A smaller specialized survey may therefore be just as, if not more, informative as a large multiple-purpose survey for evaluating the impact of particular programs or ‘treatments’ such as migration.

After trimming, the kernel matching estimates from the PINZMS and TLFS are similar, at \$329 and \$344, respectively, and close to the bias-adjusted nearest neighbor matching result of \$330. They are still 20-26% higher than the experimental estimate. Moreover, the kernel matching estimators exceed the experimental estimate in 91-94% of bootstrap replications.

It is also noticeable in Table 5 that the propensity score estimates are not very sensitive to the specification used. This differs substantially from the NSW case considered by Dehejia and Wahba (2002) and Smith and Todd (2005a,b), where the estimates are highly sensitive to the specification used. Dehejia (2005) notes that sensitivity of the matching estimator to small changes in the specification used is one diagnostic for measuring the quality of the comparison group. However, the difference in sensitivity may also reflect the relative size of the treatment effect – here the

experimental treatment effect is at least an order of magnitude larger than in the NSW program and is more significant, which may explain why small differences in the comparison groups have less effect on the results.

Thus, regardless of the matching approach used, the treatment effect is overstated by at least 20 percent when compared with the experimental benchmark. Recall that consistency of matching relies on selection on observables. But migrants are also likely to be self-selected in terms of unobservables like drive and ability, which may be more highly rewarded in the New Zealand labor market than in Tonga. Some evidence consistent with this is seen in McKenzie, Gibson and Stillman (2007). There we show that, conditional on income and employment in Tonga, individuals who apply for the PAC ballot expect to earn more income in New Zealand than do non-applicants.

5.8. Instrumental Variables with a Non-Experimental Instrument

Like with regression approaches, matching estimators only control for selection on observables, so will overstate the income gains if migrants are more talented, have more drive or have other characteristics that cause their change in income to be greater than that of observationally similar non-migrants. An alternative approach is to find a variable or variables that predict migration, but do not affect income earned in either the origin or destination countries conditional on migration status.²⁶ The advantage of an instrumental variables approach is that it explicitly recognizes that migrants have both different observable and unobservable characteristics.

We consider two potential instruments for migration in equation (3). The first is the total number of relatives an individual has in New Zealand, motivated by migration

²⁶ An example is Munshi (2003), who uses rainfall in Mexican villages as an instrument for migration when looking at the effect of migration networks on job outcomes in the United States. Given the small size of Tonga, weather variation does not provide an instrument in our application

studies that use migration networks as an instrument (e.g. Woodruff and Zenteno, 2007, McKenzie and Rapoport, 2007). This variable is significantly correlated with migration (first-stage F-statistic is 6.8). However, the exclusion restriction may be violated for this instrument in our application, since many migrants in our survey found their first job in New Zealand through relatives and hence income in New Zealand likely depends on the strength of an individual's network.

The most important reason given by non-applicants for not applying for the PAC ballot was that they did not know the requirements, which 98% of non-applicants listed as a very important reason for not applying. This motivates our choice of a second potential instrument, which is how close the individual's house in Tonga is (or was, for migrants) to the DoL office in the capital city.²⁷ We assume that distance from the office is a good proxy for information frictions since information about the requirements of the Pacific Access Category is obtained from this office, paperwork and help with the applications occurs there and the applications have to be delivered there.²⁸

It seems plausible that distance within Tonga to the DoL office in the capital city should not affect labor market outcomes in New Zealand. However, more of a concern for identification is the possibility that this distance affects income earned within Tonga. This is likely to be the case for Tonga as a whole, as individuals living on outer islands may not have the same income-earning opportunities as those on the main island of Tongatapu. However, there is only a single labor market within the island of Tongatapu,

²⁷ GPS coordinates were taken of each dwelling (former dwelling for migrants) in our Tongan survey, and of the DoL office location, and based on these, the (log) of the distance between each household and the DoL office was calculated.

²⁸ The family network in New Zealand may also be a source of information about the PAC program. This is a further argument why the network instrument may satisfy the instrument relevance condition. However, as we have argued, it seems likely that the family network variable does not satisfy the exclusion restriction.

where all villages are within one hour of the capital city. Thus, this instrument is more likely to be valid when the sample is restricted to the main island. Appendix Table 1 shows that log distance is a significant predictor of applying for the ballot, with F-statistics of 25 for Tonga as a whole, and 40 when the sample is restricted to Tongatapu.

Table 4 shows the results when using each of these variables to instrument for migration status in equation (3). Using migrant networks results in an estimated income gain from migration of \$498. This is 82% higher than the experimental estimate and exceeds the experimental estimate in 87% of bootstrap replications. There are two potential explanations for why the estimated income gains are overstated here. First, as noted above, the exclusion restriction may be violated, since a large network likely leads to higher incomes in New Zealand. Second, the instrumental variables model may be estimating a different treatment effect from the experimental estimator. If the gains from migration are heterogeneous and individuals self-select into migration based on their unobserved idiosyncratic return from it, this estimator will recover a local average treatment effect (LATE). In this case, the LATE is the income gain for (the unobserved group of) individuals who would migrate only if they had many relatives in New Zealand and who would not migrate otherwise. Regardless of which of these two explanations prevails, this approach does not produce an estimate that is close to the average treatment effect from the experiment.

In contrast, the last rows of Table 4 show that using the distance instrument results in an estimated income gain that is very close to that from the experiment, particularly when we restrict to the sample to just Tongatapu, where the exclusion restriction is more likely to hold. The estimated income gain is \$277, which is only 1.1%

higher than the experimental estimate.²⁹ Again, if returns are heterogeneous, this is only revealing the local average treatment effect. However, since 98% of non-applicants did not know about the requirements, it seems plausible that almost all of the population of interest is affected by this instrument, leading the LATE to be similar to the ATT.

5.9. Comparing the Different Non-Experimental Estimators

The analysis above has found that, with the exception of the distance instrumental variable estimator, all of the non-experimental estimators overstate the income gains from migration. Moreover, most of the non-experimental estimators give estimates larger than the experimental estimate in over 90% of bootstrap replications. The question which then arises is a relative one: which non-experimental estimators perform best?

Table 6 compares the relative performance of 11 selected estimators in each of the bootstrap replications. Consider the first column. It shows the percentage of bootstrap replications where the absolute bias of a given non-experimental estimator is smaller than that of the single difference estimator. We see that the difference-in-differences estimators, distance instrument, bias-adjusted matching estimator and kernel matching estimator all perform well relative to the single difference estimator. The second and third columns show that the majority of estimators also perform better than OLS.

The final column then calculates the percentage of replications in which a particular estimator has the lowest bias out of the 11 estimators considered here. The distance instrument performs best, with the lowest bias in 29% of replications. Difference-in-differences with polynomials performs best 20% of the time and the bias-adjusted matching estimator 15% of the time. The linear OLS specification has lowest

²⁹ The experimental estimate is \$261 when restricted to Tongatapu, so the distance instrument is within 6% of this.

bias in only 1.1% of replications. Bias-adjustment and inclusion of past income definitely improve the performance of the nearest neighbor matching.

The differences-in-differences estimators result in similar point estimates as the more sophisticated matching approaches, but with lower standard errors. The good performance of difference-in-differences in our case is thus likely a result of the PINZMS non-applicant sample already being a reasonably good match for the migrants by virtue of them being sampled in the same villages with the same questionnaire. If past income had been available in the TLFS, we would likely have found difference-in-differences to perform relatively worse on this larger, more diverse sample.

6. Conclusions

The ballot used to select migrant applicants to New Zealand from Tonga provides a unique natural experiment which can be exploited to estimate the income gains from migration and to examine how successful non-experimental methods are in estimating these gains. We estimate that there is a 263% increase in income after the first year of migrating. While this increase is large, it is only half of that suggested by differences in per capita income between the two countries and is less than that calculated using non-experimental methods to compare observationally similar migrants and non-migrants. Tongan migrants are found to be positively selected in terms of both observable characteristics and unobserved attributes such as ability and drive, both which are positively rewarded in the labour market.

The survey we use is unique in providing a natural experiment to estimate the gain from migration. In other circumstances, economists will need to rely on non-experimental estimates when making predictions and so our comparison of experimental

and non-experimental methods provides useful guidance as to the potential bias that this can entail. Our results show that a good instrument (log distance to the office where ballots are deposited) works well, but also illustrate the perils of using an instrument like migrant networks that has good first-stage power but a priori questionable excludability or which affects only a subset of the population. Among the other non-experimental methods, difference-in-differences and propensity score matching with bias-adjustment work best. However, both require collecting information on the past income of migrants, which is possible in special surveys of migrants, but not typically available when using receiving country census or labor force data.

We have also shown that GPS data can be useful in analysis, and while such data, again, are not typically collected in general labor force surveys they should be collected in specialised surveys. This highlights the importance of better data collection as a first step towards more accurate predictions of the impacts of policy changes on migration. Furthermore, like Smith and Todd (2005b), we find that the vast majority of individuals in a large labor force survey are dissimilar from the migrants, and are dropped when carrying out propensity-score analysis. Thus the smaller sample size is typical of a specialised survey may not come at much cost in terms of the number of observations available for analysis.

The estimates we obtain of the income gains from migration and our finding of positive selection on unobservables apply to the specific case of 18 to 45 year-olds migrating from Tonga to New Zealand. Nevertheless, migration in search of better employment opportunities is one of the most common forms of migration worldwide, and in many important respects, Tongan migrants are not atypical of the average developing

country migrants elsewhere in the world, suggesting that the results may apply more broadly. The average Tongan migrant in our sample has 11.7 years of education, compared to 11.0 years for the average 18-45 year-old new arrival in the United States, and much less than the 15.1 years for the average 18-45 year-old new arrival in highly skill-selective Canada.³⁰ Tongan migrants average 1.2 more years of schooling than non-migrants, a similar degree of positive selection on observables to the 0.8 years higher education of Mexican migrants moving to the United States.

References:

- Abadie, A., D. Drukker, J.L. Herr and G.W. Imbens (2001) "Implementing Matching Estimators For Average Treatment Effects in Stata", *The Stata Journal* 1(1): 1-18.
- Abadie, A. and G.W. Imbens (2006) "Large Sample Properties Of Matching Estimators For Average Treatment Effects", *Econometrica* 74(1): 431-97.
- Angrist, J.D. (2004) "Treatment Effect Heterogeneity In Theory and Practice", *Economic Journal* 502: C52-C83.
- Ashenfelter, O. (1978) "Estimating the Effects of Training Programs On Earnings", *Review of Economics and Statistics* 60: 47-57.
- Becker, S. and A. Ichino (2002) "Estimation of Average Treatment Effects Based on Propensity Scores", *Stata Journal* 2(4): 358-77.
- Borjas, George J. (1987) "Self-selection and The Earnings Of Immigrants", *American Economic Review* 77(4): 531-53.
- Borjas, George J., Stephen G. Bronars, and Stephen J. Trejo. (1992) "Self-selection and Internal Migration in the United States," *Journal of Urban Economics* 32: 159-85.
- Chiquiar, D. and G. Hanson (2005) "International Migration, Self-Selection, and the Distribution of Wages: Evidence from Mexico and the United States", *Journal of Political Economy* 113(2): 239-81.
- Deaton, A. (1997) *The Analysis of Household Surveys: A Microeconomic Approach to Development Policy*, Johns Hopkins University Press, Washington DC.
- Dehejia, R. (2005) "Practical Propensity Score Matching: A Reply To Smith and Todd", *Journal of Econometrics* 125(1-2): 355-64.
- Dehejia, R. and Wahba, S. (2002) "Propensity Score Matching Methods for Non-Experimental Causal Studies", *Review of Economics and Statistics* 84(1): 151-161.

³⁰ See Appendix 1 in the working paper (McKenzie, Gibson and Stillman 2006) for a comparison along other dimensions. The Tongan migrants we study are equally as likely to work as new migrants aged 18 to 45 in the United States and Canada, and lie somewhere between the U.S. and Canadian migrants in terms of average age, percent married, and percent female. This appendix also compares migrants to non-migrants, and Mexican new arrivals in the U.S. to non-migrants in Mexico.

- Gibson, J. and McKenzie, D. (2007) "The Impact of an Ex-Ante Job Offer Requirement on Labor Migration: The New Zealand-Tongan Experience", pp. 215-233 in C. Ozden and M. Schiff (eds.) *International Migration, Economic Development and Policy*. World Bank, Washington D.C.
- Glewwe, P., M. Kremer, S. Moulin and E. Zitzewitz (2004) "Retrospective vs. Prospective Analyses of School Inputs: The case of flip charts in Kenya", *Journal of Development Economics* 74(1): 251-268.
- Ham, J, X. Li and P. Reagan (2006) "Propensity-score matching, a Distance-based Measure of Migration, and the wages of young men", Mimeo. University of Southern California.
- Heckman, J., Hohmann, N., Smith, J. and Khoo, M. (2000) "Substitution and Dropout Bias in Social Experiments: A Study of an Influential Social Experiment", *Quarterly Journal of Economics* 115(2): 651-694.
- Heckman J., Ichimura H. and Todd P. (1997) "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme", *Review of Economic Studies* 64 (4): 605-654.
- Lalonde R (1986). "Evaluating the Econometric Evaluations of Training Programs." *American Economic Review* 76: 604-620.
- McKenzie, D., J. Gibson and S. Stillman (2006) "How Important is Selection? Experimental Vs Non-experimental Measures of the Income Gains from Migration", *World Bank Policy Research Working Paper No. 3906*.
- McKenzie, D., J. Gibson and S. Stillman (2007) "A Land of Milk and Honey with Streets Paved with Gold: Do Emigrants have Over-optimistic Expectations about Incomes Abroad?", *World Bank Policy Research Working Paper No. 4141*.
- McKenzie, D., and Rapoport, H. (2007) "Network Effects And The Dynamics Of Migration And Inequality: Theory And Evidence From Mexico", *Journal of Development Economics* 84(1): 1-24.
- Munshi, K. (2003) "Networks in the Modern Economy: Mexican Migrants in the United States Labor Market", *Quarterly Journal of Economics* 118(2): 549-597.
- Parsons, Christopher, Ronald Skeldon, Terrie Walmsley and L. Alan Winters (2005) "Quantifying the International Bilateral Movements of Migrants", Mimeo. The World Bank
- Rosenbaum, P. and Rubin, D. (1983) "The Central Role of The Propensity Score In Observational Studies For Causal Effects", *Biometrika* 70: 41-55.
- Smith J and Todd P (2005a). "Does Matching Overcome Lalonde's Critique of Nonexperimental Estimators." *Journal of Econometrics* 125 (1-2): 305-353.
- Smith J and Todd P (2005b). "Rejoinder", *Journal of Econometrics* 125 (1-2): 355-64.
- Tonga Statistics Office (2004) *Report on the Tonga Labour Force Survey 2003*. Statistics Department, Ministry of Finance, Nuku'alofa, Tonga. Available online at: <http://www.spc.int/prism/Country/TO/stats/pdfs/LFS/LFS2003.pdf> [accessed August 23, 2007].
- Walmsley, T.L. and L.A. Winters (2003) "Relaxing the Restrictions on the Temporary Movements of Natural Persons: A Simulation Analysis", *CEPR Discussion Paper No. 3719*.
- Woodruff, C. and Zenteno, R. (2007). "Remittances and Microenterprises in Mexico." *Journal of Development Economics*, 82(2): 509-28.

TABLE 1: TEST FOR RANDOMIZATION

Comparison of Ex-ante characteristics of principal applicants in successful and unsuccessful ballots

	Sample Means		T-test of equality of means p-value
	Successful Ballots	Unsuccessful Ballots	
Age	33.6	33.9	0.80
Years of schooling	11.9	11.6	0.57
Proportion male	0.55	0.51	0.58
Proportion born on Tongatapu	0.75	0.79	0.51
Proportion who are married	0.60	0.65	0.52
Height of males	175.8	173.8	0.22
Height of females	166.4	167.3	0.57
Past income (NZ dollars per week)	80.9	76.6	0.74
Proportion with the following family members living in NZ at time of last application:			
Father/Father-in-law	0.38	0.44	0.45
Mother/Mother-in-law	0.40	0.35	0.46
Brother/Brother-in-law	0.72	0.71	0.78
Sister/Sister-in-law	0.64	0.60	0.63
Aunt or Uncle	0.65	0.55	0.17
Total Sample Size	120	78	

TABLE 2: EXPERIMENTAL ESTIMATES OF THE INCOME GAIN FROM MIGRATION

Dependent Variable: Weekly Income from Work in New Zealand Dollars

	(1) OLS	(2) OLS	(3) OLS	(4) IV	(5) IV	(6) IV
Ballot Success Dummy	90.63*** (24.6)	89.74*** (24.2)	87.39*** (22.5)			
Male Dummy		-29.07 (24.4)	-23.86 (22.0)		-33.10 (23.1)	-27.77 (20.9)
Married Dummy		-4.493 (27.4)	24.54 (23.3)		-10.69 (26.6)	18.38 (22.4)
Age Dummy		0.558 (1.64)	-0.886 (1.26)		0.987 (1.54)	-0.462 (1.13)
Years of Education		13.43** (6.60)	4.605 (3.89)		12.03** (6.05)	3.274 (3.58)
Born on Tongatapu Dummy		29.17 (18.9)	27.60* (14.7)		29.59 (18.0)	28.01** (13.7)
Height		1.281* (0.65)	0.381 (0.42)		1.249** (0.61)	0.353 (0.38)
Past income			0.662*** (0.095)			0.660*** (0.090)
Migration Dummy				274.0*** (61.5)	281.1*** (61.6)	273.7*** (54.9)
Constant	104.1*** (11.8)	-297.9** (121)	-60.42 (81.5)	104.1*** (11.7)	-285.0** (116)	-48.59 (73.8)
First stage F-statistic on instrument:				66.53	61.88	61.51
Observations	197	191	190	197	191	190
R-squared	0.04	0.14	0.27			

Robust standard errors in parentheses. *, **, and *** indicate significance at the 10%, 5% and 1% levels

TABLE 3: SELF-SELECTION OF MIGRANTS

Dependent Variable: Past Weekly Income from Work in Tonga in New Zealand Dollars

	Selection into the Lottery		Selection into Migration	
	All 18-45 year olds		Lottery Winners	
	(1)	(2)	(3)	(4)
Migration Dummy Variable	55.95*** (10.6)	31.08*** (9.38)	13.14 (15.2)	5.00 (13.92)
Controls	No	Yes	No	Yes
R ²	0.15	0.32	0.01	0.15
Number of Observations	234	221	119	117

Robust standard errors in parentheses. *, **, and *** indicate significance at the 10%, 5% and 1% levels

Controls include gender, age, marital status, years of education, place of birth, height and the total number of relative types in New Zealand.

TABLE 4: NON-EXPERIMENTAL ESTIMATES

Method:	Sample Size	Estimate	s.e.	Percent Difference Compared to Experiment	Percent of Bootstrap Replications where this is greater than Experimental Estimate	Bootstrap 90% confidence interval for difference (non-experiment-experiment)
1) Using pre-migration income as the counterfactual	63	341.3	46.4	24.6	92.7	[-11.0, 150.2]
2) Selection on Observables: OLS regression						
PINZMS regressions including height						
Base specification	230	360.0	41.2	31.4	97.6	[17.9, 150.7]
Polynomials	230	347.5	42.5	26.8	95.2	[1.9, 148.5]
PINZMS regressions excluding height						
Base specification	244	362.6	40.9	32.3	98.3	[19.0, 153.7]
Polynomials	244	357.2	42.7	30.4	96.8	[11.2, 158.3]
TLFS regressions						
Base specification	4043	368.9	43.9	34.6	99.5	[28.4, 158.6]
Polynomials	4043	358.2	44.0	30.7	98.6	[15.9, 148.0]
3) Difference-in-Difference Regression						
Base specification	219	330.5	42.2	20.6	90.8	[-12.9, 120.8]
Polynomials	219	333.9	43.6	21.9	86.9	[-22.5, 130.8]
4) Instrumental Variables using migrant network	219	497.5	233.7	81.6	86.6	[-154.4, 750.0]
5) Instrumental Variables using log distance to NZIS office						
All Islands	219	309.0	89.5	12.8	67.9	[-121.6, 201.9]
Tongatapu only	159	277.1	90.0	1.1	54.3	[-165.9, 147.8]

TABLE 5: MATCHING ESTIMATES

	Sample Size	Estimate	s.e.	Percent difference compared to experiment	Percent of Bootstrap Replications where this is greater than Experimental Estimate	Bootstrap 90% confidence interval for difference (non-experiment-experiment)
Abadie and Imbens (2006) nearest-neighbor matching on multiple covariates						
<i>Single Nearest Neighbor Matching</i>						
A: Matching without using past income						
PINZMS sample using height as a control						
ATT	230 [#]	364.0	52.8	32.9	98.7	[21.0, 155.9]
bias-adjusted ATT	230 [#]	349.6	54.0	27.6	96.5	[5.4, 139.6]
Bias-adjusted ATT with interactions						
PINZMS excluding height	244 [#]	334.1	54.9	21.9	90.0	[-18.1, 136.7]
TLFS	4043 [#]	359.2	46.7	31.1	98.6	[15.5, 147.2]
B: Matching using past income						
ATT	219 [#]	345.9	56.7	26.2	96.6	[4.8, 138.9]
bias-adjusted ATT	219 [#]	334.8	58.4	22.2	91.5	[-8.8, 127.0]
C: Matching using past income and interactions						
ATT	219 [#]	342.2	59.5	24.9	96.3	[3.5, 135.8]
bias-adjusted ATT	219 [#]	330.1	59.0	20.5	85.7	[-29.2, 141.6]
<i>Multiple Nearest Neighbor Matching</i>						
D: Bias-adjusted ATT, using past income and interactions						
Nearest 2 neighbors	219 [#]	330.6	56.0	20.7	90.4	[-14.1, 123.2]
Nearest 5 neighbors	219 [#]	328.6	54.5	19.9	90.9	[-15.5, 123.0]
Nearest 10 neighbors	219 [#]	330.4	53.3	20.6	90.7	[-13.9, 122.3]
Propensity Score Matching						
<i>Radius Matching on the Common Support</i>						
Radius 0.10	156	352.3	49.5	28.6	n.a.	n.a.
Radius 0.05	150	371.5	52.7	35.6	n.a.	n.a.
<i>Kernel Matching on the Common Support</i>						
Gaussian kernel	163	329.6	47.9	20.3	92.1	[-11.2, 124.2]
Epanechnikov kernel	163	344.8	58.7	25.8	91.7	[-12.2, 142.1]
<i>Kernel Matching on Trimmed Samples</i>						
ATT trimming 0.01 and 0.99						
PINZMS	217	336.5	41.2	22.8	93.9	[-4.4, 132.0]
TLFS	1385	367.7	42.1	34.2	98.7	[21.9, 154.9]
ATT trimming 0.05 and 0.95						
PINZMS	200	329.2	49.5	20.1	93.8	[-3.2, 139.9]
TLFS	354	344.5	61.9	25.7	91.2	[-16.5, 186.3]

Note: All matching includes gender, age, marital status, years of education and place of birth. Height is also included in all of the PINZMS matches except where indicated otherwise. Interactions are interactions of sex with each covariate, quartics in age and years schooling, and an interaction between age and schooling. n.a. not available, since the Becker and Ichino radius matching program (attr.do) would not find matches on certain bootstrap replications.

Sample sizes for Nearest Neighbor Matching show the sample size over which nearest neighbors were sought. When only a single nearest neighbor is used for each migrant, the actual number of controls used is approximately equal to the number of migrants (less due to the same control being used as a match for multiple migrants, and more when there are ties in the closest match).

Table 6: Comparison of Bias Across Non-experimental Estimators

Percentage of Bootstrapped Replications where Estimator in Row has smaller absolute bias than estimator in column

<i>Non-experimental Estimator</i>	Single Difference	OLS base	OLS polynomial	DD polynomial	IV distance	Matching with past income	Bias-adjusted Matching with past income	Kernel Matched PSM	Percent where this has lowest absolute bias
Single Difference		64	54	32	44	48	35	32	9.4
OLS -base	36		29	18	37	19	16	8	1.1
OLS -polynomial	46	71		23	42	42	31	24	2.3
DD -base	72	96	85		55	84	51	54	6.6
DD - polynomial	68	82	77		57	67	52	52	19.9
IV - network	16	14	14	11		12	12	9	5.0
IV- distance in Tongatapu	56	63	58	44		56	48	45	29.4
Matching									
ATT w/o past income	34	43	31	20	37	7	14	4	1.0
ATT with past income	52	81	58	33	45		22	10	1.7
bias-adjusted ATT with past income	65	84	69	49	52	78		54	14.9
Kernel matched PSM	68	92	76	48	55	90	46		8.7

APPENDIX: NON-EXPERIMENTAL REGRESSIONS

Dependent Variable: Weekly work income for Columns 1-3 and 5-7.

Current Weekly work income - past weekly work income for column 4

COEFFICIENT	(1) OLS	(2) OLS	(3) OLS	(4) DD	(5) IV	(6) IV	(7) IV
Dummy for Migration	385.9*** (45.9)	360.0*** (41.2)	368.9*** (43.9)	330.5*** (42.2)	497.5** (234)	309.0*** (89.5)	277.1*** (90.0)
Male Dummy		76.00*** (23.1)	22.37*** (2.25)	55.48** (23.8)	74.18*** (25.8)	72.91*** (22.6)	76.67** (33.2)
Married Dummy		-31.55 (27.5)	1.933 (2.61)	-29.87 (30.2)	-23.13 (31.6)	-37.28 (30.6)	-53.31 (40.5)
Age		2.797 (1.78)	1.104*** (0.17)	0.897 (1.95)	1.876 (2.18)	2.788 (1.94)	5.025* (2.64)
Years of Education		-2.589 (2.51)	8.726*** (0.56)	-9.448*** (2.99)	-8.525 (6.81)	-3.564 (3.77)	-1.071 (4.89)
Born on Tongatapu		38.29* (21.7)	5.237** (2.33)	18.29 (23.1)	19.59 (32.0)	38.59 (25.1)	37.99 (39.6)
Height		3.589** (1.40)		4.468*** (1.49)	3.351** (1.52)	4.164*** (1.38)	4.885** (1.98)
Past Income					-0.0796 (0.43)	0.252 (0.29)	0.171 (0.30)
Constant	41.91*** (4.22)	-664.3*** (215)	-111.3*** (9.23)	-697.2*** (224)	-556.5** (268)	-742.9*** (220)	-933.7*** (318)
Instrument Islands	All	All	All	All	Network All	Distance All	Distance Tongatapu
First stage F-stat:					6.84	24.94	39.56
Control sample:	PINZMS	PINZMS	TLFS	PINZMS	PINZMS	PINZMS	PINZMS
Observations	230	230	4043	219	219	219	159
R-squared	0.45	0.51	0.37	0.43			

Robust standard errors in parentheses. *, **, and *** indicate significance at the 10%, 5% and 1% levels