Economic Mobility in South Africa: Evidence from Household Survey Data

Arden Finn

Thesis presented for the Degree of

DOCTOR OF PHILOSOPHY

in the School of Economics

Faculty of Commerce

UNIVERSITY OF CAPE TOWN

February 2017

Abstract

High levels of inequality, poverty and unemployment are some of the most substantial challenges facing post-apartheid South Africa. Most of the research addressing these questions has used micro datasets to compare snapshots of welfare over time. Although these studies are both interesting and useful, they have been unable to extend their analysis into a nationally-representative dynamic setting, due to the lack of available data. The paucity of large longitudinal datasets has also limited the number of studies of economic mobility, which allows researchers to track the welfare measures of the same individuals over time. This means that while we know a great deal about how South Africans are doing at a particular point in time, we know far less about how they are faring dynamically. Understanding how and why economic mobility happens in South Africa is therefore a question that demands attention. From both a distributive justice as well as a policy point of view, the distinction that arises when we drop the assumption of anonymity and move from a cross-sectional measure of welfare to a dynamic one is important. This is because many of the conclusions about longer-run welfare are dependent on the level of economic mobility present in society.

This study contributes to the body of work on welfare in South Africa by addressing three different aspects of economic mobility. The first of these is about how a particular kind of measurement error in household surveys is best detected, and what effect its presence has on the understanding of labour market mobility. The second is about how best to model money-metric poverty dynamics in South Africa in order to better understand who escapes poverty and who enters poverty over time. The third is about how the persistence of intergenerational earnings should be calculated in a society with high unemployment, and what the role of education is in shaping these mobility dynamics.

Acknowledgements

I am deeply grateful to Murray Leibbrandt and Vimal Ranchhod who have provided excellent guidance along the path of my research in general, and for this thesis in particular. My colleagues at the Southern Africa Labour and Development Research Unit (SALDRU) have created a wonderful environment for collaborative, relevant and productive research.

I gratefully acknowledge generous financial support from the National Research Foundation's (NRF) Social Dynamics in Development Grand Challenge, NRF funding through the Chair in Poverty and Inequality Research, the Programme to Support Pro-poor Policy Development (PSPPD), and the Research Project on Employment, Income Distribution and Inclusive Growth (REDI). I also thank everyone on the team at the National Income Dynamics Study (NIDS) who were a pleasure to work with throughout the production of four very interesting waves of data.

Last and most I would like to thank Gabi for her endless reserves of patience and unwavering support as I commandeered dining room tables in lieu of office space in various cities around the world.

Table of contents

List of figures

List of tables

1 Introduction

1.1 Motivation for this thesis

The central questions that motivate this thesis are, "Why are some people so poor, why are some people so wealthy, and what should be done about it in the interests of advancing a just society?" This is particularly apposite in a South African context, where centuries of legislated structural poverty, inequality of opportunities and outcomes, and state-endorsed discrimination were ended - in a political sense at least - in 1994. A great deal of recent research has shown that the economic gains made by the majority of South Africans have been slow to converge on the political gains made in the country since the end of apartheid.

Measuring progress requires data, and to this end there have been a multitude of snapshots which compare South African poverty and inequality over time using post-apartheid micro datasets.[1] These studies, while both interesting and useful, have been unable to extend analysis into a nationally representative dynamic setting due to the lack of available data. The paucity of large longitudinal datasets has also limited the number of studies of mobility - tracking welfare measures of the same individuals over time. This means that while we know a lot about how South Africans are doing at a particular point in time, we know far less about how they fare dynamically. Understanding how and why economic mobility happens in South Africa is therefore a question that demands attention.

From the points of view of both distributive justice and policy, the distinction that arises when we drop the assumption of anonymity and move from a cross-sectional measure of welfare to a dynamic one is important. This is mainly because conclusions about longer-run welfare are dependent on the level of economic mobility present in a society. For example, it is possible to obtain identical measures of cross-sectional welfare, poverty and inequality at two points in time, even if nobody's welfare remained the same in absolute or relative terms. Once we drop the assumption of anonymity and follow the same people over time, we are able to ask, and answer, different questions.

There is a poignant motivation for studying how welfare[2] changes intertemporally and intergenerationally. In South Africa's high-inequality society the degree of mobility over time may

[1] Micro in this context means data from individual, household and labour market surveys.

[2] Welfare in this context may refer to household income, household consumption expenditure, labour market status, and labour market earnings, among other things.

impact both the demand for redistribution and social stability, a point made in a South African context by Pellicer et al. (2014), and more generally by Friedman (1962). Indeed, many of the recent protests in South Africa that have taken place around tertiary education and service delivery have been fuelled largely by participants' concerns about inequality. Higher mobility (even if it is higher perceived mobility rather than actual mobility) may be coupled with a higher tolerance for inequality, as a relative deficit today could easily be reversed tomorrow. On the other hand, the median voter theorem implies that low mobility and persistently high inequality will lead to stronger demands for redistribution, as well as a more fragmented and fractious society. Determining who gains, who falls behind and who remains trapped over time is therefore an important undertaking.

In a broad sense this thesis is about economic mobility in South Africa. More specifically, it takes three different departure points on economic mobility and evaluates questions related to these. Part of the reason for adopting different viewpoints of economic mobility is because, as a concept, it is rather more fluid than poverty or inequality, and there is less consensus on what exactly economic mobility is. When poverty and inequality are measured there is generally a single concept in mind, with different indices used to derive a deeper understanding of the research question. When economic mobility is the subject, however, it is common not only to talk of different indices, but of different concepts entirely. As noted by Gary Fields, "Not only do different people have different ideas about what economic mobility is, but they have different clear ideas about what economic mobility is." Fields (2010), (emphasis in original).

1.2 Structure of the thesis

The three main chapters of this thesis address distinct yet related aspects of economic mobility in South Africa. In short, the first is about measurement error and mobility, the second is about absolute mobility as understood through poverty dynamics,[3] and the third is about intergenerational earnings mobility and the role played by education in determining the persistence of

[3] Absolute poverty in this chapter means using absolute poverty lines when analysing poverty dynamics, rather than using a relative poverty line such as a fraction of mean or median income. It also means that ranking individuals in the income distribution is not important, because the focus is where people are relative to a poverty line, rather than relative to one another.

earnings over time.

These chapters ask a number of distinct questions about economic mobility in South Africa. Chapter 2's investigation asks the following:

- What impact does measurement error from interviewers have on measures of economic mobility?
- What is the most effective method for identifying this kind of measurement error?
- How does this error feed through to cross-sectional versus longitudinal estimators?

Chapter 3 asks:

- Who is escaping poverty, who is entering poverty, who is staying in poverty and why?
- What are the relative roles of income and demographic events in determining mobility across the poverty line?
- How can we model poverty dynamics while accounting for initial conditions, endogeneity and selective attrition?

The central questions in Chapter 4 are:

- How should the persistence of earnings between parents and children be calculated when there is high unemployment in both generations?
- What is the shape of the intergenerational correlation of earnings across the wage distribution?
- What is the role of education in shaping these mobility dynamics?

Mobility, when understood as an empirical concept, needs to be measured after it has been defined. Chapter 2 addresses the dual problems of how to detect a certain kind of measurement error in survey data, and what effect that error has on how we understand economic mobility. Of particular interest is a certain kind of measurement error - fieldworker fabrication of data - and how this feeds into the evaluation of categorical mobility. This is applied to labour market transitions and the movement (or lack of movement) into and out of employment. The focus is on non-classical measurement error, whereby an individual is assigned an incorrect labour market status. The impact of measurement error on labour market status variables (and categorical variables in general) in longitudinal data can be particularly problematic, as the following simple example demonstrates. Suppose person M is employed in wave 1 and wave 2 of a panel dataset. Now suppose that person M is incorrectly classified in the data as being unemployed in wave 1, and correctly classified as being employed in wave 2. This kind of measurement error creates not only a cross-sectional misclassification (in wave 1), but the false impression of labour market mobility (wave 1 to wave 2). As this chapter shows, there may be situations in which this kind of misclassification is intentional, and this can have consequences for our understanding of mobility over time. As will be demonstrated, this kind of measurement error may have muted effects when on simple descriptive statistics, or even OLS regressions. However, as researchers begin to exploit the time series properties of the data more and more, the detrimental presence of this measurement error becomes increasingly apparent.

Chapter 3 retains the theme of absolute mobility, but shifts the focus from the labour market to poverty. It does this by evaluating some of the reasons for why so many South Africans have been unable, to paraphrase Carter and May (2001), to see the concomitant benefits of economic freedom that should have come with political freedom in 1994.[4] In doing so the chapter expands the South African literature from a province-specific lens[5] to a national perspective. To this end, four waves of longitudinal household survey data from 2008 to 2015 are used to present the descriptive statistics of poverty transition and persistence in South Africa. Following this, the chapter assesses the relative importance of income events versus demographic events in determining poverty transitions over time. Finally, the dynamics of poverty in the country are estimated using an endogenous switching model that takes special account of the changing composition of our longitudinal data, and the role of initial conditions in determining longer-run poverty dynamics. This has important policy implications, as the nature of poverty, along with the relative shares of chronic versus transitory poverty, can help shape how we tackle the

[4] Carter and May (2001) themselves base the title of their article on Ransom and Sutch (1977).

[5] Carter and May (2001), Cichello et al. (2005) , Woolard and Klasen (2005) and Agüero et al. (2007) all use data from KwaZulu-Natal.

problem in the future.

In Chapter 4 the focus shifts to earnings mobility across generations. Measuring the level of persistence in the earnings of one generation to the next can be a challenging undertaking in a developing country context. This is further complicated if unemployment rates are very high, as they have been in South Africa over the last two decades. This chapter's initial question is about how the correlation of earnings between parents and children should be calculated given conditions of high unemployment. It then moves on to investigate the role of education in shaping this relationship. Intergenerational dynamics are investigated using four waves of household survey data from 2008 to 2015 along with an additional nationally representative dataset from 1993. As will be discussed in depth, using both sets of data allows for the correction of possible biases that arise from co-resident parent-child pairs, and from selection into labour market participation in a high-unemployment society like South Africa's. The presence of intergenerational poverty and inequality traps is investigated through using quantile regression techniques to estimate persistence across the range of earnings, and a similar approach is taken in order to determine the importance of educational attainment in driving the intergenerational correlation of earnings across the wage distribution.

Finally, Chapter 5 provides a summary of the key results of this thesis.

2 Genuine fakes: The prevalence and implications of data fabrication in a large South African survey

2.1 Introduction

For anyone involved in the running of a survey, issues of data quality are of critical importance. Surveys can cost millions of dollars, require years of planning by large teams of people and need considerable levels of sustained effort. All of these resources are allocated for the sole purpose of producing high-quality data. All empirical findings, in turn, are premised on the assumption that the data being used are of a reasonable quality. This caveat applies to vast literatures in economics, sociology, demography and political science, amongst others. Indeed, it is so ubiquitous that it hardly ever gets stated explicitly.

In this chapter, we investigate one aspect of the data production process that might lead us to question the quality of survey data in general, and its effect on the measurement of dynamics in particular. Most survey organizations, either directly or indirectly, employ interviewers to conduct their surveys. The interviewers, though, might not have the same objectives as the survey organization. These principal-agent problems might result in interviewers 'cheating'.[1]

Interviewers may engage in cheating behaviour for a variety of reasons. First, interviewers may be reluctant to ask sensitive questions about topics related to income, wealth or sexual behaviour. Second, some sections are very long and interviewers might want to leave them out in order to save time. Third, the characteristics of the primary sampling unit (PSU) may play a role. If the PSU is in an area that is considered dangerous (which is not uncommon in the South African context) or is very far away, then interviewers may end up cheating rather than visiting the PSU. Fourth, interviewers might be remunerated according to the number of successful interviews that they have completed. In the case of refusals, or the case where it is easier to fabricate an interview, this would incentivize cheating. Finally, the penalties for cheating may be small. If the survey company is unable or unwilling to monitor the behaviour of the interviewers, then the expected payoff to cheating might exceed the expected costs for some workers.

There are also different ways in which cheating behaviour could manifest itself. First, and

[1] We use the word 'cheating' although in some cases a better word might be 'negligence'. The former implies intent whereas the latter could arise out of ignorance, incompetence or misunderstandings, and we cannot always distinguish between the two. In either case, interviewers do something that they ought not to have done which results in a deterioration of the aggregate quality of the data produced.

most problematic, interviewers could fabricate entire interviews. In later waves of longitudinal studies, there is usually some pre-population of the questionnaires based on data from previous waves. This often includes a list of household members from the roster and their demographic characteristics. Interviewers can view this information and use it to form the basis of their fabrication. Interviewers could also cheat by leaving out sections of interviews. For example, in wave 1 of the National Income Dynamics Study (NIDS) questionnaire, which is the longitudinal dataset that we use in this study, the labour market section is substantial and has a total of 89 questions.[2] However, a respondent who is 'not economically active' will only answer seven simple 'yes/no' questions. Interviewers could save time by setting respondents' labour market statuses to 'not economically active' when they are, in fact, working or looking for work. A different way to save time would be to leave out certain people in the household. This would be easy to implement in a cross-sectional study. In a longitudinal study, the interviewer might ignore new members in the household, such as babies or in-migrants, or exaggerate the number of people from the previous wave who have died. Our research findings presented in this chapter are primarily concerned with the most problematic type of cheating listed, namely the fabrication of entire interviews. We find that analysis of cross-sectional numerical patterns and longitudinal anthropometric measures are the most effective means of detecting data fabrication in our context. Our methods suggest that approximately 7% of the sample was affected in this way. If the fabrication had not been detected, it would not have substantially affected our cross-sectional estimates, but would have led us to reach different findings as more complex, longitudinal, estimators are used. In particular, we would have systematically overstated transitions into the 'not economically active' labour market category, and systematically underestimated transitions into employment and unemployment from any wave 1 base state. Furthermore, the presence of fabricated data would have led us to reach substantially different conclusions about the role of labour market mobility as a determinant of changes in BMI across waves 1 and 2. A brief cost-benefit analysis of the data quality investigation suggests that the benefit was more than 24 times the aggregate cost.

The remainder of this chapter is structured as follows. In section 2.2, we argue that the inci-

[2]This includes 10 sub-questions.

dence of interviewer cheating is a common problem in the implementation of large household surveys in several countries, including South Africa. In section 2.3 we turn our focus to the first two waves of the NIDS dataset and evaluate a number of methods that we considered to detect interviewer cheating.[3] The two most successful methods, Benford's law and anthropometric diagnostics are dealt with in greater detail than the others. Section 2.4 analyses what the consequences for future research would have been, had the cheating not been detected and corrected for, and compares the benefits of the data quality investigation to the aggregate costs of detecting fabricated data. Section 2.5 offers recommendations for future fieldwork operations and provides some concluding remarks.

2.2 The prevalence of interviewer cheating in survey data

The phenomenon of interviewers making up data is a global and persistent problem with a sizeable literature dedicated to documenting and detecting it. A number of studies use data from major surveys in the United States to detect whether data fabrication had occurred. Schreiner et al. (1988) use Census Bureau Studies data from 1982 to 1987 to highlight the importance of reinterviewing respondents as a means of fraud detection. In their study, 83% of suspected falsifications turned out to indeed be a result of cheating. Most of the cheating was detected through reinterviews, although some were picked up because of anomalies in the data. In addition, most of the cheating involved total rather than partial fabrication of individual-level data. The authors find that falsification rates range from 0.4% to 6.5% depending on which one of the Census Bureau surveys is used.[4] Finally, they note that interviewers who had served for longer periods of time are significantly less likely to be data fabricators. Li et al. (2011) make the point that the Census Bureau's reinterview strategy for detecting falsification can be improved upon. The conventional reinterview methods detect falsification in less than 0.1% of the data. The authors use data from the Current Population Survey to try to design an alternative

[3] This work was done by the author while wave 2 of NIDS was still in the field. At the time, the author was employed in the NIDS office.

[4] Fraud prevalence rates in this and future examples refer to the proportion of individual, rather than household, interviews that contained falsified data.

sampling method that should underlie the reinterview process. Using a combination of real data and simulations, they conclude that alternative sampling methods could find up to 20% more fabricated interviews than the current system. Murphy et al. (2004) use data anomalies from the National Survey on Drug Use and Health to identify suspicious interviewer behaviour. In particular, they flag relatively short or long interview durations as possible signs of falsification and show how taking these durations into account adds to the power of the fraud detection process.

Outside of the US, Schäfer et al. (2004) use data from the German Socio-economic Panel (SOEP) to test the reliability of two methods of fraud detection. Data fabrication was low in all waves and all samples of the SOEP, never exceeding 2.4% of all cases, with the overall share of faked data at about 0.5% (Schräpler and Wagner, 2005). The authors use the fabricated data that was removed from the publicly-released version of the SOEP and find that using Benford's law as the basis for detecting suspicious data correctly identifies all cases of fabrication. In addition, they exploit the fact that cheating interviewers tend to have less variability in their responses over all questions and all interviews than non-cheaters. Interviewer-level tests for surprisingly low variance also correctly identified all of the cases of cheating. All confirmed cheating interviewers were middle-aged and male, and the effect of education on the probability of cheating was not statistically significant.

There are a number of other studies that use characteristics in the data themselves as a means of identifying fabrication. These include, amongst others, Bredl et al. (2008) in an unspecified non-OECD country, and Porras and English (2004), Cho et al. (2003) and Swanson et al. (2003) in the US. A broad review on much of the literature related to the detection of fabricated data can be found in Birnbaum (2012) who charts the methods used in twelve different datasets in the developed and developing world.

Although the fabrication of survey data is an issue of concern to researchers throughout the world, the remainder of this study narrows the focus somewhat by drawing attention to illustrative cases of interviewer cheating in the South African context. These examples highlight the existence, but not the prevalence, of data fabrication in South African household surveys. They motivate our study by showing that there is enough prior evidence to take this phenomenon seriously, even though they are agnostic as to how problematic or widespread this is in household survey datasets.

2.2.1 KwaZulu-Natal Income Dynamics Study (KIDS)

KIDS is a household level panel dataset that was conducted in 1993, 1998 and 2004. It revisited a subset of the households located in the KwaZulu-Natal province of South Africa that were included in the original SALDRU/PSLSD 1993 survey. Follow-up fieldwork in May of 2001 suggested that there may have been cheating by interviewers in some clusters. Subsequent investigations revealed that the fabrication was limited to two clusters and these were permanently removed from the sample.[5] Judge and Schechter (2009) compare data from the deleted clusters to data from the retained clusters, and find large differences between the two in the module on crop production and animal ownership.

2.2.2 Survey on time and risk preferences

Between 2010 and 2011, researchers from the University of Cape Town conducted a survey on time and risk preferences in the three major metropolitan regions of South Africa,[6] with a budget of about 300 000 US dollars.[7] They had a sample size of about 300 respondents and visited each of them six times at three monthly intervals. The survey included a background questionnaire as well as two experimental modules. In the experimental modules, respondents were asked to choose between various alternatives in order to ascertain their appetite for risk and their discount factors. In order to obtain truthful responses, all choices were incentivized to have some probability of entailing an actual cash payout.

In the time preferences component, respondents answered 40 questions. They then rolled a 10-sided die, and if it landed on 0, they would get paid for one of their 40 responses. The relevant question was selected by rolling a 10-sided die and a 4-sided die simultaneously. In the risk preferences component, respondents were also asked 40 questions, one of which would

[5] See May et al. (2007) with an earlier version available for download at `http://www.datafirst.uct.ac.za/catalogue3/index.php/catalog/286`.

[6] These are Johannesburg/Pretoria, Cape Town and Durban.

[7] Information on the details of this study was obtained through interviews with Andre Hofmeyr. At the time, he was a researcher actively involved in the survey.

yield a cash payout with certainty. The relevant question was also selected by means of rolling a 10-sided die and a 4-sided die simultaneously. The payouts varied by question and by the choices made by the respondents in that question. The interviewer would then pay the amount of the winnings in cash to the respondents.

After the data were collected, researchers found a suspiciously high rate of interviewees getting paid out for the time preferences component. The *ex ante* probability of this occurring was 10% but the respondents were 'winning' 25% of the time overall. Moreover, respondents were observed to have a disproportionately high probability of having 'randomly selected' questions with relatively higher cash payouts in both the risk preferences and time preferences component of the study. Further investigation indicated that these anomalies were driven by data from a subset of interviewers who almost always paid out the maximum amounts permissible. People involved in the study believe that some interviewers colluded with respondents so as maximize the actual disbursements, which they could then share. The problem was identified only after the 4th wave of data had been obtained, with the 5th wave already in the field, and both the time preferences and risk preferences components of the study had to be abandoned.

2.2.3 Cape Area Panel Study: Wave 5

The Cape Area Panel Study (CAPS) is a longitudinal study of young adults in the Cape Town metropolitan area. Wave 1 was conducted in 2002 with a sample of about 4 800 young adults aged 14 to 22. In the fifth wave of CAPS, conducted in 2009, part of the interview included a finger-prick test for HIV status which was administered by the interviewer.[8] The *ex ante* expectation was that about 30% of the women interviewed would be HIV positive. For most interviewers, the proportion of HIV-positive female respondents was indeed reasonably close to 30%, but after a certain date, one interviewer returned HIV-positive results for every respondent.

It took a considerable amount of the time for the lab results on the blood to be returned to the operational headquarters. Thus, by the time that this was discovered, the interviewer in question

[8]The information on CAPS was obtained through interviews with Brendan Maughan-Brown, co-ordinator of the fifth wave of CAPS. More information can be obtained in Lam et al. (2012) which can be downloaded at `http://www.datafirst.uct.ac.za/catalogue3/index.php/catalog/266`.

had already been paid and had left the survey. Investigations discovered that the interviewer in question had not, in fact, taken blood samples from respondents, but had obtained blood samples from some other source. The result was that all data collected by this interviewer after a certain date was deleted and did not form part of the fifth wave.

A common method of monitoring interviewer behaviour is to phone respondents in the weeks or months after the interview in order to verify that they were in fact interviewed. One of the interviewers obtained the list of verification questions and set up a system in which her sister-in-law pretended to be a respondent each time she was called by the survey company. This suggests that interviewers who do cheat can use quite sophisticated methods to avoid detection. This interviewer's cheating was only discovered after the conclusion of fieldwork, and all relevant data was deleted from the study.

Overall, a total of 8 interviewers had engaged in some form of cheating, out of an average of about 40 interviewers over the course of the fieldwork. A total of 289 fraudulent interviews were deleted from the public release version, which represented about 9% of the expected sample at the start of wave 5.

2.2.4 Time Use Study: 2000

In 2000, StatsSA, the official statistical agency of South Africa, conducted a national time use study over three different months in order to investigate how South Africans spend their day. The total sample size was approximately 14 000. Household members were eligible to participate if they were aged 10 or older. Interviewers were asked to fill in a household roster in descending order of age, and if more than 2 household members were eligible, to select two household members sequentially using a sampling grid.[9] If the interviewer reached the end of the grid, that is at the 11^{th} such household, then she was instructed to 'loop' back to the start - that is, to treat the 11^{th} household with 4 eligible people as if it were the 1^{st} household with 4 eligible people that she had encountered (Statistics South Africa, 2000).

The sampling grid yields an asymptotic distribution of the frequency with which household

[9] A copy of the sampling grid is included in Appendix 2.A as Table 2.A.1. To illustrate how it works, suppose that an interviewer came to her first house with four eligible members. She should then select persons 2 and 4, i.e. the 2^{nd} and 4^{th} oldest members of the household. In the second such household, she should select persons 1 and 3, etc.

members of a particular age-rank ought to have been selected on aggregate, conditional on the number of eligible persons. For example, in households with three eligible persons, we would expect to find that person 1 was selected 50% of the time, person 2 was selected 70% of the time and person 3 was selected 80% of the time.[10] In the dataset, however, we find that persons 1, 2 and 3 in households with three eligible persons were in fact selected 81%, 81% and 38% of the time, respectively. A Chi-squared test rejects the null hypothesis that the realized distribution corresponds to the expected distribution at any reasonable level of significance.[11] We repeated the analysis for households with 4, 5 and 6 eligible members respectively, and convincingly rejected the null hypothesis of equivalence of distributions in each case. We interpret this as evidence that interviewers did not, in fact, follow instructions about whom to select in households where they had some degree of choice.

An alternative explanation to interviewer cheating is that the asymptotic distribution is not the correct distribution to use, since we would only expect it to be realized if each interviewer had 'many' households with more than two eligible persons. Nonetheless, at least in the case with three eligible persons, there are more than 1 000 such households in total. If each interviewer encountered 10 or more such households then the asymptotic distribution would provide a reasonable approximation of the correct expected distribution. Moreover, the magnitude of the divergence is so great that, unless the expected distribution that we use is grossly incorrect, we would continue to reject the null hypothesis that the realized distribution corresponds to the 'true' expected distribution.

Note that in this case we are not claiming evidence of data fabrication. The 'cheating' that we observe here is of a very different nature to those previously documented, and it is important to draw the distinction between intentional and unintentional sources of data error. What explains such cheating? We conjecture that interviewers violated the sampling instructions due to some combination of the availability of respondents as well as variations in the time that it would take for different respondents to complete the questionnaire.

Two empirical observations support this conjecture. First, in households where the interviewer had some choice, 53.7% of those eligible were female, but 56% of those selected were

[10]The total adds up to 200% since two household members were selected.
[11]The results of this analysis are presented in Table 2.A.2 in Appendix 2.A.

female. This difference is small in terms of percentage points but since it applies to just over 9 000 observations it is statistically significant. Moreover, there is nothing in the sampling grid that suggests a clear gender bias in terms of who ought to be selected. A more plausible explanation is that females are more likely to be available for an interview, since they are much less likely to be employed in South Africa.[12] Second, we observe that in households with 3 or more eligible persons, 51% of those eligible are younger than 21 years of age. Of these potential respondents, only 35% were selected to fill in a questionnaire. Teenagers are probably less likely to be available due to being in school. Additionally, they might be less willing to participate in interviews in the first place, and it might take longer for them to complete a questionnaire.

If our conjecture is correct, then the violation of the sampling framework has potentially serious implications for analyses. Obtaining a disproportionate number of unemployed or 'not economically active' people in the sample will bias measures of aggregate time use, and conventional sampling weights, even if adjusted for non-response, will not correct this bias.

2.2.5 Labour Force Survey 2001

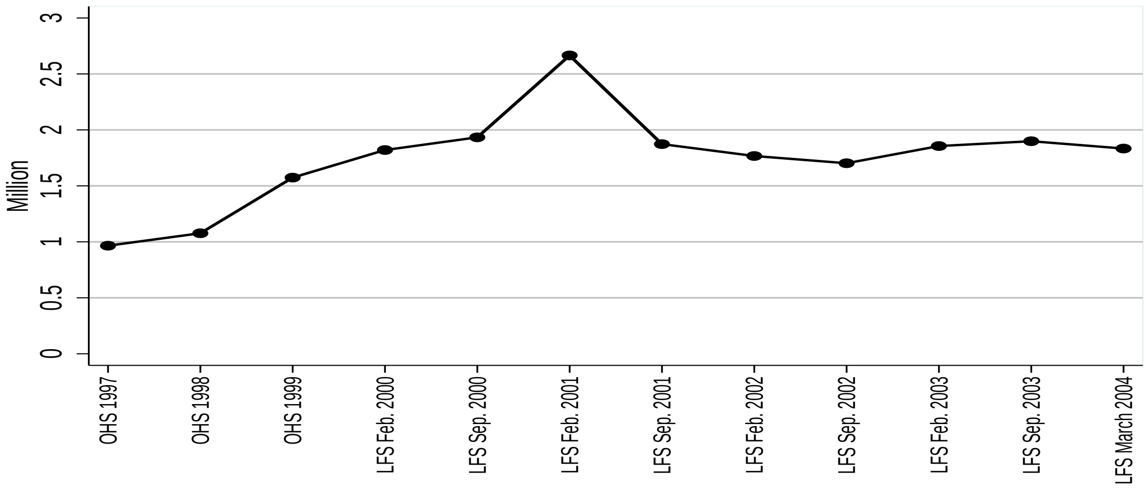
Devey et al. (2006) contains an interesting figure, reproduced below as Figure 2.1, that charts the number of people in South Africa classified as 'informally employed'. The authors use data from October Household Surveys (OHS) and Labour Force Surveys (LFS) from 1997 to 2004. The most striking feature of the data is the spike in February 2001, with a jump of almost 750 000 informal workers, before falling by approximately the same number to September 2001. It is implausible that such a spike should be present in nationally representative data with a consistent survey instrument. The authors point out that in the February 2001 LFS, interviewers were offered additional incentives to interview informal workers in the households that they had visited.

Although there is no established claim for cheating having taken place whereby interviewers fabricated data on the informal economy, it is nevertheless very suspicious to see the spike when finding informal sector workers was incentivized, and an immediate and equivalent reversal

[12]In StatsSA's September 2000 Labour Force Survey (LFS), amongst prime working-age adults aged 21 to 59, the employment rates for men and women were 61.6% and 45.5% respectively. This was a nationally representative survey with a sample size of about 50 000 working-age adults.

when the incentive was removed. In addition, we cannot be certain as to whether the February 2001 data is incorrect, or whether all of the other data is reflecting a systematic under-counting of informal employment, or both. The example nonetheless highlights that data quality is potentially affected in a substantial way by fieldworker incentives.

Figure 2.1: Informal employment (in millions) from OHSs and LFSs 1997 – 2004



Source: Reproduced from Devey et al. (2006).

In summation, we have provided evidence of interviewer cheating in four substantial South African surveys and highlighted potential cheating in a fifth. These surveys have been both cross-sectional and longitudinal, and span a time period from 1993 to 2011. The implementing agencies include local and international fieldwork companies, as well as organisations that employed and managed interviewers directly. The cheating manifests in various forms, including outright fabrication of entire interviews, falsification of responses to particular questions and not following the sampling instructions. In some cases the cheating did not affect the overall legitimacy of the study, whereas in one case an entire component of the study needed to be abandoned. The research areas that have potentially been affected include time use, health, risk preferences, labour market status, poverty, inequality and intergenerational mobility. Thus the incidence of interviewer cheating is widespread and its effects are potentially far reaching.

In the next section, we discuss how we attempted to identify possible cheating in wave 2 of NIDS.

2.3 Interviewer cheating in wave 2 of the National Income Dynamics Study

The second wave of NIDS is the main focus of this chapter. NIDS is a nationally representative longitudinal study that collects data from respondents on many socio-economic topics including education, labour market participation, fertility, mortality, migration, income, expenditure and anthropometric measures. The survey starts with a household roster that documents all people who are resident in the household at the time of the interview. From the information captured in the roster, respondents are classified as either children (aged 0–14) or adults (aged 15 and above). Interviewers are instructed to then administer either a ‘child questionnaire’ or an ‘adult questionnaire’ for each household resident. In cases where the respondent refused or was not available, interviewers were asked to try to get a knowledgeable person in the household to fill in a ‘proxy questionnaire’ on behalf of the respondent who could not be interviewed.

The first wave, which took place in 2008, had a sample size of 7 301 households and about 17 000 people completed the adult questionnaire. In that wave, interviewers used paper-based questionnaires and entered responses by hand. The completed questionnaires were then sent from the fieldwork company to a data capturing company and, by the time the full dataset was received by the survey operations team, fieldwork had already been completed. The primary data quality control procedures thus occurred after the fieldwork had been completed in wave 1.

The second wave of NIDS was conducted over 2010 and 2011. Interviewers changed from the paper-based questionnaires of wave 1 to a Computer Assisted Personal Interview (CAPI) system, whereby interviewers filled in responses on a hand-held computer. Data from completed questionnaires were then uploaded to a server on a daily basis. One of the advantages of having data come in ‘live’ was that a verification process was undertaken while the interviewers were still in the field. This allowed for corrections to be made as part of the ongoing fieldwork operations, so that suspicious data could be verified or replaced, rather than deleted.

Our objective for the verification process was to create a measure that could rank interviewers by decreasing levels of suspiciousness. Once interviewers were ranked, the respondents that they interviewed were called back to ascertain whether they had been interviewed or not, and if they had indeed been interviewed, whether the entire questionnaire had been completed.

In creating the suspicion-based ranking of interviewers we considered using nine different methods. The central idea in using each of these possibilities was that interviewers who do cheat will do so either to save time, or to earn more money, or both. Interviewers could earn more money as they received a performance-based incentive for each completed individual and household questionnaire. This would result in systematic differences in some dimensions of the data that were generated by cheating interviewers, when compared either to the data obtained from non-cheating interviewers or to externally-motivated benchmarks. The usefulness of a method was determined by its ability to identify anomalies in the data, and to rank interviewers by the prevalence of these anomalies. The subsequent success (or lack thereof) of the method in identifying fraudulent data was ascertained by callbacks to households which were recorded as having been visited by each suspicious interviewer.

The most successful of these nine methods were the use of Benford's law and anthropometric comparisons, which we discuss in detail in sections 2.3.2 and 2.3.3 respectively. Although the other seven methods were not particularly useful for diagnostic purposes, we document what did not work as it might still be useful to people running surveys in other contexts.

2.3.1 Unsuccessful methods of detection

Method 1: Number of deaths between waves

One way to speed up the process of completing an entire household would be to falsely classify a household member from wave 1 as deceased. This would allow a fieldworker to 'complete' interviewing the household much faster. Alternatively, fieldworkers could falsely classify someone who had died between waves as being still alive, and then fabricate the data. We compared the mortality rates of respondents by fieldworker, and did not observe any anomalies in the data.

Method 2: Number of refusals/not available

The method and thought behind using this metric is identical to that for using deaths above. Fieldworkers could save time by not interviewing everyone (for example, by fabricating refusals and non-responses from respondents), or fabricate data for those who had, in fact, refused to be interviewed. We compared the response rates by fieldworker, and did not observe any anomalies in the data.

Method 3: Fieldworkers who are disproportionately likely to activate substantial skip patterns in the survey

Our thoughts here were that one could save considerable time in some sections by capturing certain responses. As already discussed, this incentive is strongest in the labour market section. We abandoned this method as the levels of unemployment in South Africa are high, levels of labour force participation are low, and these are concentrated in certain neighbourhoods and regions (Leibbrandt et al., 2010). Since fieldwork was co-ordinated geographically, fieldworkers could plausibly have genuinely encountered a pool of respondents with low levels of employment and labour force attachment in their allotted region. In addition, the unemployment rates and percentage that were not economically active, by fieldworker, yielded several fieldworkers with high values, so this was not a particularly useful tool for discriminating between suspicious and non-suspicious fieldworkers.

Method 4: Using length of interviews to identify fabrication

If fieldworkers were fabricating data, we expected them to complete the surveys relatively quickly. The software we used had time stamps for both when the interview began and was completed, which theoretically allowed us to calculate the time per interview. We also expected that each adult interview would take between 45 minutes and one hour to complete. Unfortunately, the time stamp for completion was activated manually, and several fieldworkers only did so at night prior to uploading the data to the server. This rendered this component of the investigation useless.

Method 5: Using GPS co-ordinates to verify where the interview took place

Part of the survey captures the GPS co-ordinates of the household. This was required for all households in both wave 1 and wave 2. The co-ordinates were obtained by means of a handheld GPS device which was accurate to a radius of 100m. If interviews were being fabricated, we would expect to find differences between the wave 1 and wave 2 co-ordinates. We encountered two problems with this method. First, there was considerable measurement error in wave 1, so not all differences could be attributed to wave 2 cheating. Second, fieldworkers were given the GPS co-ordinates from wave 1 to assist them in finding the households. Instead of entering the GPS readings from the GPS device in wave 2, a cheating fieldworker could simply re-enter the co-ordinates that they had received.

Method 6: Comparing wave 1 and wave 2 signatures

Each completed questionnaire in each wave needed to be accompanied by a signed paper-based consent form. We considered comparing wave 1 and wave 2 signatures to identify discrepancies. We abandoned this approach very quickly, as signatures have some variability over time, and the method was far too labour intensive.

Method 7: Low rates of in-migration or births between waves

If fieldworkers were fabricating entire households, then they would not be able to know about any new household members that entered between waves. They would then either systematically under-estimate the number of new members, or else have to fabricate these new members as well. We calculated the number of new members per household by fieldworker, but there were no clear patterns or anomalies. If cheating fieldworkers did indeed fabricate new members as well, or if cheating fieldworkers cheated only on some fraction of their households, then it would be much harder for this diagnostic to yield usable information.

2.3.2 Using Benford's law

In contrast to the methods used above, the use of Benford's law as a ranking mechanism for suspicious interviewers proved to be very useful. Following a paper by Schäfer et al. (2004), we used Benford's law as the basis of a test of the distribution of the numerical data reported by each interviewer. Benford's law is an empirical law that was first described in Benford (1938). It describes the probability distribution of leading digits in tables of numerical data and asserts that the distribution is not uniform, as might be expected *a priori*, but rather follows a certain logarithmic probability distribution given by:

$$Pr(\text{leading digit} = d) = log_{10}\left(1 + \frac{1}{d}\right), \quad d = 1, 2, ..., 9$$

This implies that the probability of a leading digit being 1 is about 30%, the probability of it being 2 is about 17.6%, with the corresponding probabilities of the subsequent leading digits decreasing monotonically until we find that the probability of the leading digit being 9 is approximately 4.6%. The probability distribution of leading digits is shown in Table 2.1, below.

Table 2.1: Benford's distribution of leading digits

1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
30.1%	17.6%	12.5%	9.7%	7.9%	6.7%	5.8%	5.1%	4.6%

For a long time this phenomenon was viewed as not much more than a numerical curiosity. However, some practical implications began to emerge (Scott and Fasli, 2001; Durtschi et al., 2004), and Benford's law has since been used to detect fraud in financial statements of companies (Carslaw, 1988; Thomas, 1989). More recently, it has been used in a wide variety of settings in the US (Durtschi et al., 2004). Judge and Schechter (2009) used Benford's law to compare the data from the deleted and retained clusters in the KIDS example discussed earlier. The law has also been found to hold with a large number of other kinds of data, including the population of towns, the length of rivers and the half-life of radioactive atoms. The basic premise is this: If you have a relatively large dataset and you accept that Benford's law holds, then you can identify possible cheating by comparing the realized distribution of leading digits for each interviewer, to the distribution of leading digits that would be expected if Benford's law holds.

Hill (1995) provided the first theoretical basis for the law and showed that the law applies most accurately to stock market data, some accounting data and census statistics. The intuition underlying the proof is the following: Consider a variable which grows at some constant rate. Regardless of the initial value or the growth rate, the asymptotic distribution of leading digits of this variable (over time) will conform to Benford's law.[13] Thus, a random sample of such variables at a moment in time will also conform to Benford's law.

More recently, Schäfer et al. (2004) and Schräpler (2011) argue that certain survey data also conform to Benford's law, and use this for the express purpose of identifying cheating interviewers in the German Socio-economic Panel (SOEP). Schräpler (2011) summarises three requirements that need to be satisfied in order for Benford's law to be a useful diagnostic for detecting fraud in survey data.[14] First, the data should not have a built-in maximum value. Second, there should be no externally assigned values in data. For example, the South African old age pension is a rand amount that is assigned to an individual, and this is an example of data that cannot be used in the diagnostic test. Finally, the distribution of the data should be positively skewed with a median that is lower than the mean. Of all the variables in wave 2 of NIDS, a number of those in the income and expenditure modules satisfied all of these criteria. Some of these are reported at the household level (for example household non-food expenditure), some are reported at the individual level (for example monthly wages), and others are aggregated across respondents in the same household (for example household wages). Variables that included any systematic component, such as the value of government grants, were not valid candidates for inclusion.

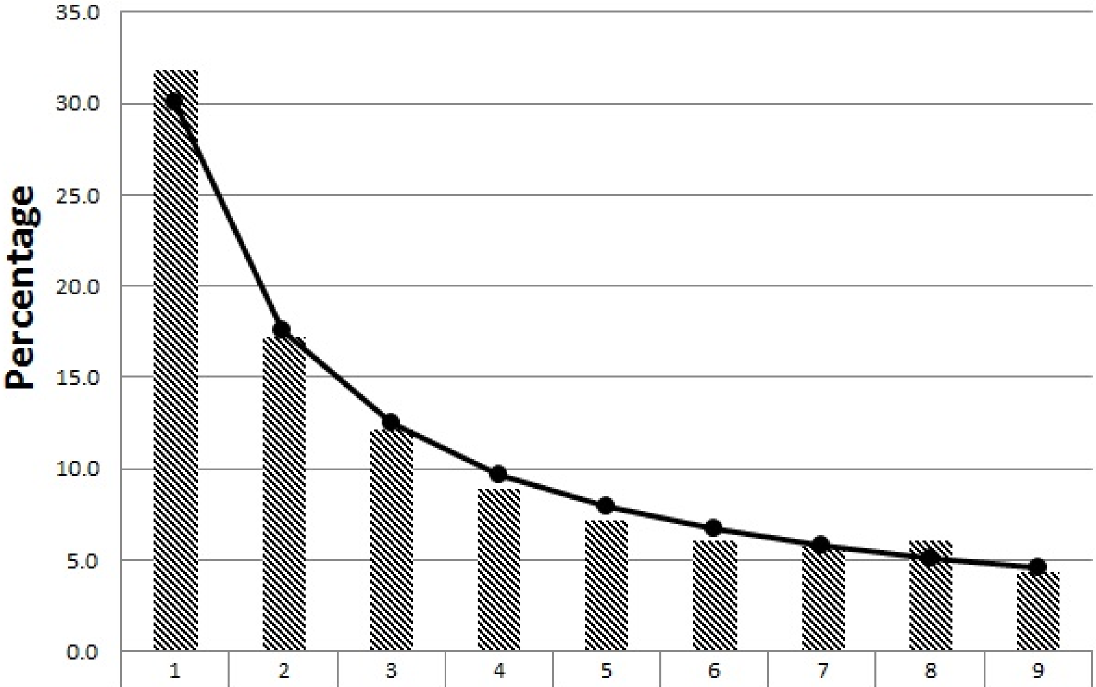
Figure 2.2, below, plots the distribution of leading digits of the variable reflecting the total amount of labour market income received by respondents in the 4 705 households with positive wages in the wave 1 data. This distribution, shown by the bars, is plotted together with the logarithmic distribution that we expect to observe, assuming that Benford's law holds. By comparing the two distributions, it seems that the leading digits of this variable fit the logarith-

[13] Informally, consider a variable that grows at a constant rate with an initial value of 1 (note that the initial value does not have to be 1, it is used here for illustrative purposes). As this number grows, more time will be spent between 1 and 2, than between 2 and 3, and more time will be spent between 2 and 3 than between 3 and 4. This amount of time with each successive leading digit decreases until the value reaches 10, in which case more time is spent between 10 and 20 than between 20 and 30, and so on. As the time period increases, the distribution of leading digits converges to Benford's distribution.

[14] The first two are validated through their presence in practical applications (see, for example Nigrini (1999)), while the third is a requirement that is derived from simulated results in Scott and Fasli (2001).

mic distribution very well. The observed proportions of each leading digit are very close to the proportions that we expect to observe *ex ante*, and fall with each successive digit, except for eight which is slightly higher than expected.[15]

Figure 2.2: Observed and expected leading digits - Wages in wave 1



	1	2	3	4	5	6	7	8	9
Observed	31.9	17.2	12.2	9.0	7.3	6.1	5.8	6.1	4.3
Expected	30.1	17.6	12.5	9.7	7.9	6.7	5.8	5.1	4.6

Source: Own calculations using NIDS Wave 1 2008.

Given that the data seem to follow this logarithmic distribution for some of the monetary variables, we next sorted the wave 2 observations by interviewer and considered the conditional leading digits of total household income as recorded by the interviewer. We ranked how far each interviewer's distribution of leading digits was from the logarithmic distribution, by computing Chi-squared statistics. Ranking interviewers by this method yielded positive results for the detection of cheating. Of the interviewers with the five highest Chi-squared values, four were subsequently found to have fabricated entire questionnaires. The top ten Chi-squared rankings are shown in Table 2.2, below. The cheating interviewers are highlighted in bold.[16]

[15]See Figure 2.B.1 in Appendix 2.B for a comparison of observed versus expected distributions of leading digits for some other wave 1 and wave 2 monetary variables.

[16]The full table for all interviewers who collected data from at least 40 households can be found in Table 2.C.1.

Table 2.2: Most suspicious interviewers by Chi-squared ranking

Ranking	**Interviewer**	**Chi-squared (no. of interviews)**
1	**A**	**39.7 (80)**
2	**B**	**31.2 (49)**
3	OK	28.0 (66)
4	**C**	**27.3 (74)**
5	**D**	**27.1 (100)**
6	**E**	**24.7 (42)**
7	OK	21.7 (64)
8	**F**	**21.4 (73)**
9	OK	21.2 (44)
10	OK	19.8 (67)

Source: Own calculations using pre-public release NIDS Wave 2 data, 2010.
Interviewers who were found to have fabricated data are identified by letter, and are in bold typeface.
Interviewers who were suspicious but did not fabricate data are denoted 'OK' and are not in bold typeface.
Number of interviews refers to the total number of household interviews submitted by the interviewer.

The fact that six of the top ten most suspicious interviewers were identified using Benford's law suggests that using this method is appropriate for survey data of this nature. Nonetheless, some cheating interviewers may have left the monetary variables blank, or set them to missing. In this case we would have the unique problem of missing fake data, which also happens to be fake missing data. One of the ways of overcoming this problem is to use repeated observations of the same respondents over time to pick up possible fabrication. We thus exploited the longitudinal dimension of the data and evaluated variables that are difficult to fabricate convincingly in a panel study – namely, height and weight.

2.3.3 Anthropometric measures

One of the advantages of a longitudinal dataset is that prior or future waves may be used to calibrate data quality in other waves. Slow-moving variables such as anthropometric measures are good candidates for this exercise. The first two waves of NIDS included modules in which respondents were weighed and measured for height. Weights were obtained using digital scales that were accurate to 0.1kg, while heights were obtained using a portable stadiometer. These data were not pre-populated into the CAPI system, making it almost impossible for interviewers to systematically fabricate values that were consistent with wave 1 measures for respondents that they had not seen.

Various diagnostic measures were used to rank interviewers according to their likelihood of having cheated. These were:

- **Mean adult body mass index (BMI), by interviewer**.[17] Our thinking was that interviewers who were fabricating data might not be aware of the extent to which the height and weight of respondents are correlated. This would result in 'abnormal' BMI measures. We considered interviewers who generated exceptionally high or exceptionally low average BMI values as potentially suspicious.

- **The mean growth in the height of adult respondents between waves, by interviewer**. We expected that the heights of adults would be stable over a two year span, on average. If the mean growth in height for a particular interviewer differed substantially from zero in either direction then we interpreted this as an indication of possible cheating.

- **The mean BMI growth from wave 1 to wave 2, by interviewer.** If this differed substantially from zero then this was a sign of possible fabrication.

- **Spikeplots of the weight distribution, by interviewer**. Given that the scales were digital, and that no interviewer had interviewed hundreds of respondents, we expected to obtain a uniform distribution of weights. Visual inspection of the spikeplots allowed us to relatively quickly identify suspicious patterns such as 'heaping' at natural reference numbers.

The diagnostics above were restricted to adults only, where adults were classified as respondents 20 years old and above. Running the diagnostics on children would have presented a problem as the height and weight variables for children are more volatile, even in a two to three year period.

Table 2.3, below, shows the list of suspicious interviewers generated using the mean adult BMI in the wave 2 cross-section. As before, interviewers who were found to have cheated are highlighted in bold. Interviewer E, who interviewed 97 adults, had the highest mean BMI of

[17] BMI is calculated by dividing a person's mass (in kilograms) by the square of their height (in meters). A BMI above 25 is considered to be overweight by the medical profession.

55.3. The highest mean BMI measure that we verified *ex post* was 43.3, but was based on only 20 respondents. At the other end of the distribution, interviewer H returned a very low mean BMI of 21.7, much lower than the overall average of 28.6 from 9 821 adults.

Table 2.3: Suspicious interviewers and mean adult BMI

Interviewer	N	Mean BMI
E	**97**	**55.3**
F	**24**	**49.6**
G	**49**	**48.9**
B	**104**	**44.7**
OK	20	43.4
OK	33	38.8
H	**156**	**21.7**
OK	62	21.5
Total	9 821	28.6

Source: Own calculations using pre-public release NIDS Wave 2 data, 2010.
Interviewers who were found to have fabricated data are identified by letter, and are in bold typeface. Interviewers who were suspicious but did not fabricate data are denoted 'OK' and are not in bold typeface. N is the number of completed questionnaires submitted containing adult BMI data.

Many of the same interviewers also appeared to be suspicious when BMI growth, rather than the mean of BMI, was used for identifying cheating. As shown in Table 2.4, six of the 12 most suspicious interviewers were subsequently found to have fabricated either part of an interview or the whole interview, for at least one of their interviews. The mean percentage change in BMI for the entire adult sample was 9%. Interviewer G, who only interviewed 32 adults, returned a BMI growth rate of 173%, followed by interviewers E and B with 109% and 99% respectively. At the other end of the distribution, H's 117 respondents showed a decrease in BMI of 19%, on average.

Table 2.4: Suspicious interviewers and adult BMI growth

Interviewer	N	Mean % change
G	**32**	**173**
E	**67**	**109**
B	**75**	**99**
OK	38	33
I	**83**	**31**
OK	89	25
OK	35	23
OK	14	20
J	**80**	**20**
OK	44	-7
OK	40	-15
H	**117**	**-19**
Total	5 560	9

Source: Own calculations using pre-public release NIDS Wave 2 data, 2010.
Interviewers who were found to have fabricated data are identified by letter, and are in bold typeface. Interviewers who were suspicious but did not fabricate data are denoted 'OK' and are not in bold typeface. N is the number of completed questionnaires submitted containing adult BMI data.

We present the suspicious list obtained by using mean adult height growth between waves in Table 2.5, below. Interviewers who were subsequently found to have fabricated data are shown in bold once again. Of the 5 710 adults for whom valid data were recorded in both waves, the mean change in height was a rise of 0.11%. Interviewers H and A recorded the largest mean growth rates of around 5%. The four most suspicious interviewers at the other end of the distribution, Interviewers E, I, B and G recorded very large negative growth in height ranging from -4.95% to -14.70%.

Table 2.5: Suspicious interviewers and mean change in adult height

Interviewer	N	Mean % change
H	**118**	**5.21**
A	**120**	**4.86**
OK	35	4.81
OK	41	4.72
OK	45	4.62
E	**68**	**-4.95**
I	**83**	**-5.38**
B	**75**	**-7.14**
G	**32**	**-14.70**
Total	5 710	0.11

Source: Own calculations using pre-public release NIDS Wave 2 data, 2010.
Interviewers who were found to have fabricated data are identified by letter, and are in bold typeface. Interviewers who were suspicious but did not fabricate data are denoted 'OK' and are not in bold typeface. N is the number of completed questionnaires submitted containing adult height data.

The final anthropometric method used to detect suspicious interviewers was a visual inspection of spikeplots of the weight distributions. This allowed us to quickly observe heaping at focal points. Moreover, one weakness of the three methods used above was that they would only diagnose cheating if the proportion of interviews that were faked was 'substantial' enough to affect the mean.[18] This method was not dependent on the mean of the weight distribution obtained by the interviewer.[19] It could thus be informative even in cases where an interviewer had cheated on only a small fraction of surveys.

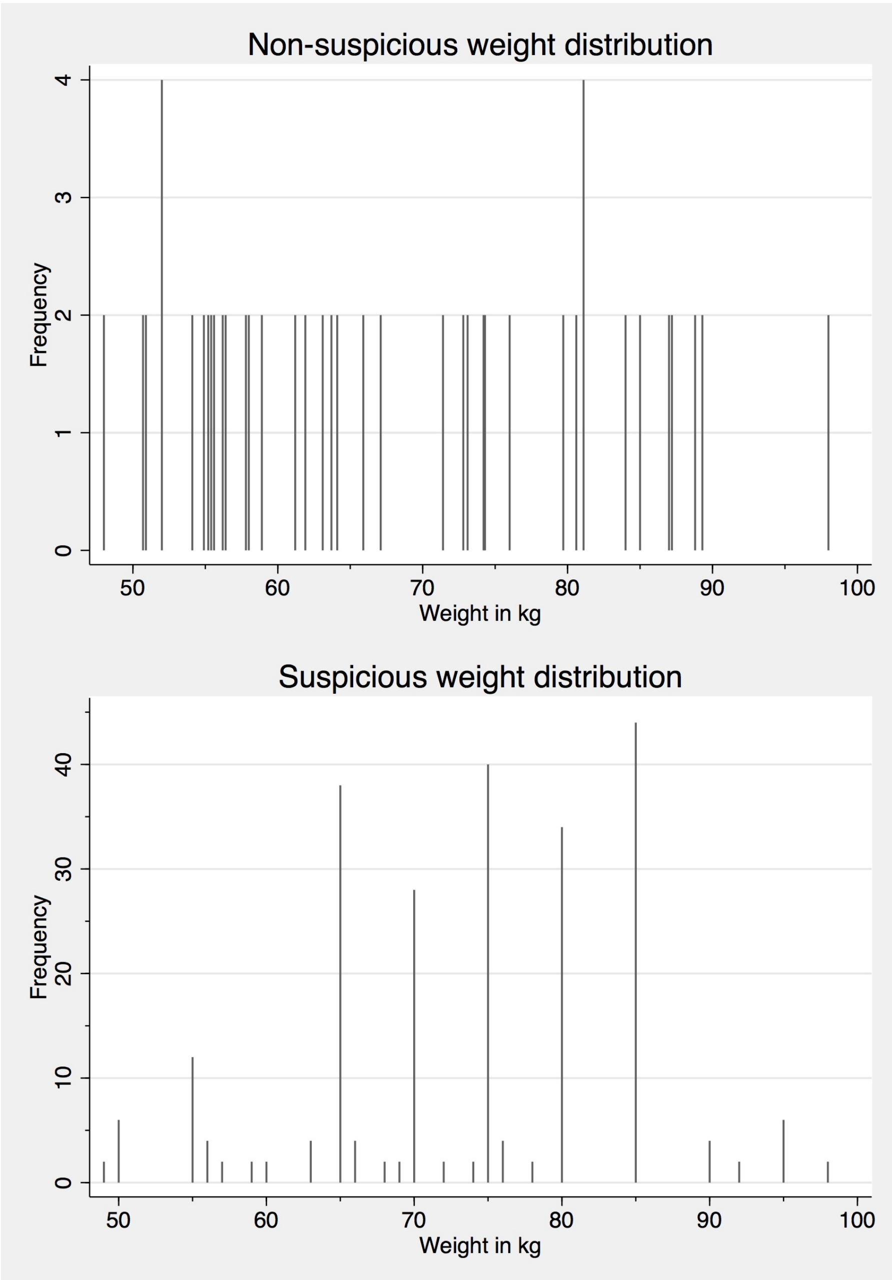
We provide two spikeplots as illustrative examples of recorded weights for adults, with values restricted between 48 and 100. The upper panel of Figure 2.3 shows the spikeplot of the weight distribution of a non-suspicious interviewer. This interviewer interviewed 39 adults (with weights between 48kg and 100kg) and recorded two weight values for each of them, hence the uniformity at a frequency of two on the y-axis. Only two adults out of the 39 had the same weight; 52kg and 81.1kg. Contrast this to the spikeplot of a suspicious interviewer,

[18] We also considered interviewers who captured anthropometric data that consisted of a 'high' number of outliers. This did not prove to be effective. Almost all interviewers had some outliers, which could arise because the respondent really did have an exceptional height or weight, or due to measurement error in either wave 1 or wave 2.

[19] An additional approach that we used was to sort interviewers by the 10^{th}, 25^{th}, 50^{th}, 75^{th} and 90^{th} percentiles of their obtained distributions of height, weight, BMI and the growth of these variables. These did not yield substantially new insights beyond those obtained from the methods already described and employed.

shown in the lower panel of Figure 2.3. This interviewer interviewed 125 adults and there is significant spiking at 65kg, 70kg, 75kg, 80kg and 85kg with relatively few of the other observations having different values. Note that the y-axis goes up to 44, compared with the upper panel where it only goes up to 4. This distribution immediately raised suspicions that the interviewer had made up the anthropometric data at best, and had fabricated the entire interview at worst.

Figure 2.3: Spike-plot of weight distributions



Source: Own calculations using pre-public release NIDS Wave 2 data, 2010.

Bringing the three anthropometric measures together allowed us to create a crude index of suspicion. Interviewers were assigned scores of zero to three, depending on how many times they were flagged as suspicious in each of the diagnostic methods. Table 2.6 shows the top 12 most suspicious interviewers according to the three anthropometric diagnostics. Every interviewer who scored three out of three was later found to have fabricated data. Of the interviewers who were flagged as suspicious using the anthropometric measures, interviewers A and B were also flagged as the two most suspicious interviewers using the Benford's law method. Constructing a joint index of the Benford scores and the anthropometric scores did not prove fruitful, as some of the highly-suspicious interviewers according to the Benford method did not provide much anthropometric data - they tended to record refusals in this section - while, in general, the suspicious interviewers according to anthropometrics failed to provide many data points for income and wage variables.

Table 2.6: Combined anthropometric suspicion index

Interviewer	**BMI**	**BMI Growth**	**Height Growth**	**Row Total**
H	**1**	**1**	**1**	**3**
E	**1**	**1**	**1**	**3**
B	**1**	**1**	**1**	**3**
G	**1**	**1**	**1**	**3**
I		**1**	**1**	**2**
OK	1	1		2
F	**1**			**1**
A			**1**	**1**
J		**1**		**1**
OK		1		1
OK			1	1
OK	1			1
Col. Total	7	8	7	21

Source: Own calculations using pre-public release NIDS Wave 2 data, 2010.
Interviewers who were found to have fabricated data are identified by letter, and are in bold typeface.
Interviewers who were suspicious but did not fabricate data are denoted 'OK' and are not in bold typeface.

An important issue to consider is that data in wave 1 could have been fabricated as well. If interviews (or parts of interviews) were faked in the first wave of data, this could feed through and make large changes in anthropometric data look suspicious. This would be problematic as the error is entering in the first period, rather than in the second period, which was the focus of our investigation. On average, however, even if the first wave contains fabricated data, this should be diluted as different interviewers interviewed different respondents in both waves. The probability that the majority of a wave 2 interviewer's valid interviews are combined with mostly fake wave 1 data is small but non-trivial, given the spatial logistics under which fieldwork was conducted in both waves. Our assumption, therefore, is not that the wave 1 data are perfect, but rather that there is no perfect overlap between respondents who were interviewed by cheating interviewers in waves 1 and 2.

One implication of the possibility that we are identifying wave 1 cheating instead of wave 2 cheating, is that we needed to be more cautious about our conclusions. The diagnostics that we employed are inherently probabilistic, not deterministic, and this lack of determinism is exacerbated by any wave 1 cheating that occurred. This provided an important motivation for the second stage of our auditing process, which we discuss below.

2.3.4 The NIDS operational response

A meta-list of suspicious fieldworkers was drawn up using a combination of the Benford's law rankings and the anthropometrics rankings. The NIDS operations centre initiated an intensive set of telephonic callbacks in order to verify whether or not the interviews of suspicious fieldworkers had actually been conducted. Priority was given to calling back respondents who were interviewed by the most suspicious fieldworkers and the NIDS team worked down the list systematically, calling every household for which data had been collected by that fieldworker, until there was a high level of confidence about the veracity of the data.[20]

The callbacks comprised a set of questions intended to establish initially whether an interview had taken place or not, and whether the entire interview had been completed. A copy of the list of questions asked in the verification process can be found in Figure 2.D.1 in Appendix 2.D.

Where fraud was evident, all data from those interviews were rejected at the expense of the company conducting the fieldwork. This was true regardless of whether there was partial or total fabrication of the relevant interview. During the data collection phase, fieldworkers were

[20]Additional random callbacks to households interviewed by non-suspicious fieldworkers were also undertaken, and these did not result in any additional fraud being uncovered.

sent out in teams of three, with a team leader and two additional interviewers. The intensive auditing revealed that data fabrication was generally a team-specific phenomenon. That is to say, if one fieldworker in Team A was found to have cheated, there was a fairly high probability that the other two team members also cheated. In all cases in which the team leader was found to have cheated, the other two fieldworkers also fabricated their data, to a greater or lesser extent. The auditing turned up one interesting case where a suspicious fieldworker was found to have used the scale and measuring device incorrectly for the anthropometrics. In this case, only the anthropometric data was flagged as invalid and the fieldworker was sent for additional training.

Once the NIDS callback team had reached the point where fraud was no longer being uncovered, a new phase of data collection was put in place to re-interview the appropriate respondents. Instead of deleting data that was fabricated, thereby reducing the sample size in the second wave, correctly collected information was re-integrated into the dataset. Figure 2.4 describes the different outcomes of the verification process. Overall, we identified 991 households that needed verification. Of these, 781 households were successfully contacted, meaning that over 10% of the total sample was called back in the verification process.[21] Of these 781 households, 234 were verified as having been validly interviewed. Furthermore, it was found that 547 households had data that was not valid, either because of partial or total fabrication. These 547 households were made up of 223 partial fabrications, 322 total fabrications and 2 which could not be classified.[22]

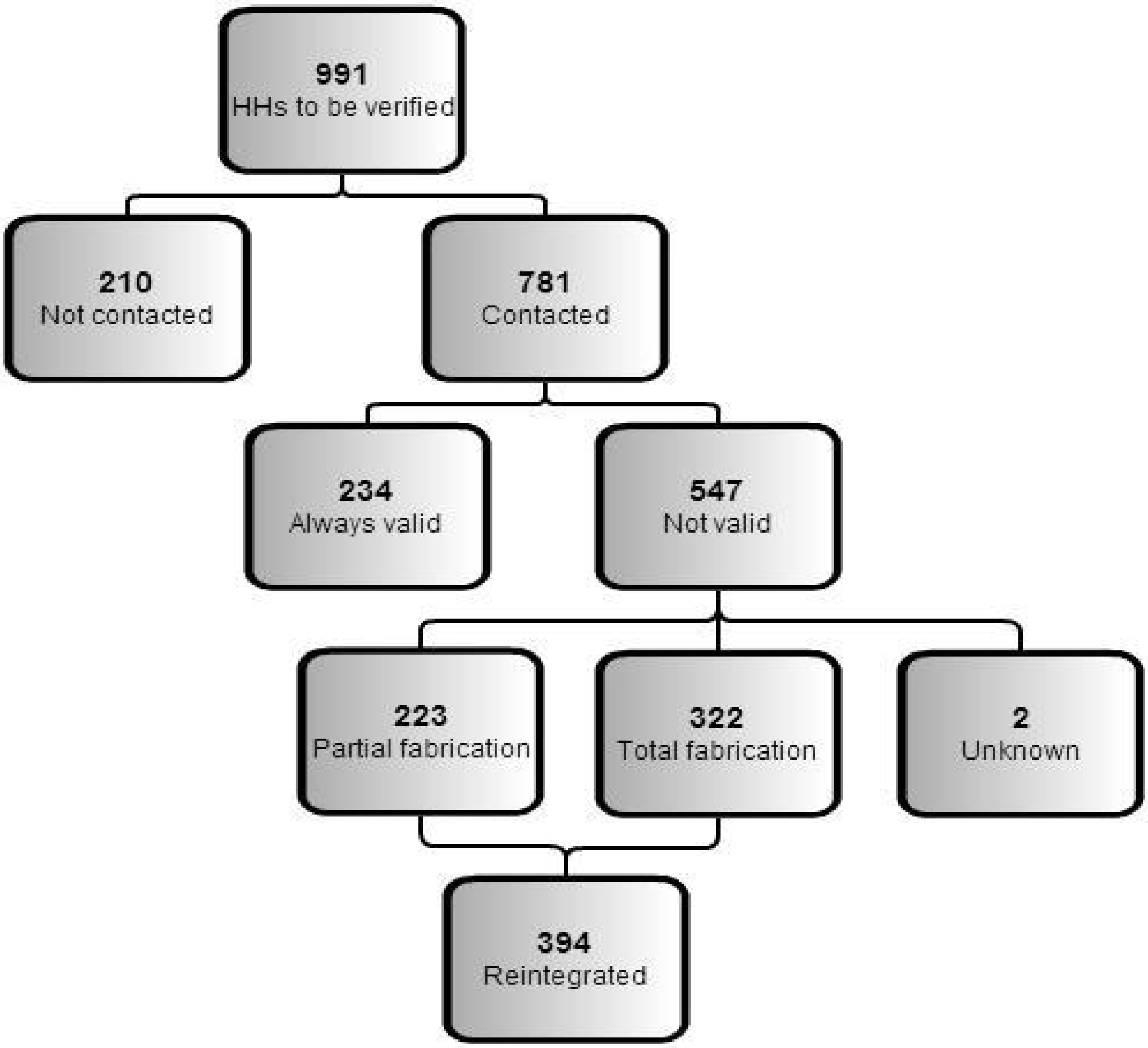
In summation, of all the households on the suspicion list that were successfully contacted, over 70% had large data quality concerns that were driven entirely by fieldworker cheating. This represents 7.3% of the wave 2 households at the time that the verification process started. Of the 547 households with problematic data, a total of 394 households were successfully re-interviewed and re-integrated into the dataset. Thirteen fieldworkers, or approximately 10% of the total number of fieldworkers that were employed to conduct fieldwork for wave 2 of

[21]The NIDS team was unable to incorporate 210 households into the verification process either because invalid phone numbers were provided and respondents could not be contacted, or because respondents refused to take part in the verification process.

[22]These two households were successfully contacted and the reference person indicated that there was a problem with the interview process, but refused to provide further information.

NIDS, had produced some data that was entirely fabricated.[23] The rates of fabrication across fieldworkers who were found to have cheated ranged from 10% to 67% of all households interviewed.

Figure 2.4: Verification process and outcomes



Source: Own calculations using pre-public release and public release NIDS wave 1 and wave 2 data, 2010/2011.

2.4 Implications for analysis

By how much would the presence of the fabricated data have affected our estimates, had the cheating interviewers not been discovered? Assuming that some fabrication is probably present

[23]Note that not all of their data was fabricated, but that some positive proportion of the data that they generated certainly was fabricated.

in most surveys, and simultaneously, that most interviewers are probably honest, should we be wary of most empirical results? Alternatively, does the measurement error caused by interviewer cheating have relatively small effects on our estimates, such that, for practical purposes, we may ignore its implications with respect to research findings? In addition to the resources invested in the production of data, considerable time, energy and resources are invested by users of these data, and research findings subsequently feed into important policy making discussions and debates. Measuring the effects of interviewer cheating on the validity of empirical findings is the objective of this section of the chapter.[24]

From an econometric perspective, data fabrication leads to measurement error for potentially all of the variables in some subset of the data. *A priori*, we cannot make a general prediction about the effects of data fabrication on subsequent estimates, as the effects, if any, will depend on multiple factors. These factors include the fraction of the overall dataset that is fabricated, the difference between the fabricated data and the true data that it represents, the type of estimator being implemented, whether the fabrication results in classical[25] or non-classical[26] measurement error in the variable or variables that are being used, and the magnitudes of such measurement error. Moreover, if one is using a multivariate estimator, the empirical effects will depend on the relationship between the variables being used in the fabricated dataset, relative to the true relationship between those variables. Any theoretical predictions thus need to be restricted by quite a specific set of criteria.

Nonetheless, there are some well known and fairly general effects that measurement error in a regressor will have in a regression analysis. First, measurement error in an independent variable will result in a violation of the orthogonality condition. This will induce biased estimates

[24]Note that this is a different kind of question about measurement error and mobility to the one asked in Burger et al. (2016), in which measurement error of continuous variables such as household income poses a serious concern for researchers of mobility who use South African household survey data.

[25]Assuming the true model is $y^* = X^*\beta + \epsilon$ but we measure $X = X^* + \mu$ and $y = y^* + v$, under the conditions of classical measurement error, u and v are i.i.d. and uncorrelated with X^*, y^* and ϵ and the estimated β coefficients are biased in the direction of zero. See Bound et al. (2001) for a comprehensive overview of the literature on bias due to measurement error.

[26]Of course, the misclassification of a categorical variable such as labour market status or a dummy variable such as employed/unemployed cannot be thought of in the same way as classical measurement error, as the error itself cannot be mean zero. In fact, for dummy variables, the measurement error must be negatively correlated with the true value of the variable. The case for measurement error in categorical variables is not as straightforward, but a thorough treatment is beyond the scope of this chapter. See Krueger and Summers (1988) for a discussion of the results of measurement error in categorical regressions.

of the $\hat{\beta}$ vector obtained from an OLS regression (Wooldridge, 2002). In the case of classical measurement error, this will result in an attenuation bias, that is, a bias of the estimated coefficient towards zero. Second, some common estimators, such as fixed effects estimators and first difference estimators, are more sensitive to a particular endogeneity problem than a standard OLS estimator (Griliches and Hausman, 1986).

In addition, Schnell (1991) and Schräpler and Wagner (2005) find that univariate statistics such as means, medians and variance are generally robust to the presence of fake data, where the prevalence of fake data is less than 5%. However, the negative effects of fake data begin to compound as analysis moves to a multivariate setting, particularly when some of the commonly-used panel data estimators are used (Schnell, 1991; Schräpler and Wagner, 2005). We should mention again that our focus in this study is data that were totally fabricated by interviewers, and not questionnaires that were partially correct and partially fake. As such, our findings should be interpreted as lower bounds of the problem of cheating in this dataset.[27]

In order to provide an illustrative example of the effects of cheating in our dataset, we chose to investigate the broad theme of understanding the effects of finding employment on health, as measured by BMI. We chose this area of investigation for two reasons. First, we have spent some time documenting the fabrication that took place in the labour market module as well as the height and weight measurement module. The extensive discussion of the weight, height and BMI variables in Section 2.3.3 provides context for the analysis that follows. Looking at the effect of finding employment on a measure of well-being such as BMI complements our section on detecting fabrication, and speaks directly to the labour market transitions matrices that follow. Second, the determinants of BMI as well as the effect of employment on BMI are topics that have received a great deal of attention in the recent South African literature (Wittenberg, 2013, 2009; Ardington and Gasealahwe, 2012; Ardington and Case, 2009) and our study provides an important addition to these.[28] Indeed, some of this research uses NIDS data to investigate the relationship between labour market status and BMI, and so the implications of fabricated data for analysis are clear beyond this chapter. We then needed to choose a

[27] We use the term 'lower bound' in a non-technical sense. That is, we mean that the extent of fieldworker fraud that was uncovered is almost surely less than the full extent of fieldworker fraud in the data.

[28] We also modelled the effect of receiving the state old age pension on labour force participation using the Dirty and Clean datasets. The results of these analyses are available from the authors on request.

set of variables and a set of estimators for our analysis. For the variables, we include BMI, age, education and labour market status. Following the theoretical discussion in the preceding paragraphs, we calculate the mean BMI, the labour market transition rates and finally, we fit OLS and First Difference regression models of BMI on age, education and employment.[29]

2.4.1 Data

To implement the analysis, we constructed two datasets. The first, which we refer to as the 'Dirty' dataset, is a combination of the 'Always Correct' data combined with the 'Fake' data at the time that our verification process was completed. Essentially, it represents what the NIDS wave 2 dataset would have been if the cheating had gone undetected, and the survey was completed at the date that our verification process drew to a close. The second dataset, which we refer to as the 'Clean' dataset, is composed of the same 'Always Correct' data, combined with the subsequently corrected data where such correction was possible.

The variables that we use are all at the individual level. They are:

- **BMI** - This is calculated as a person's mass in kilograms divided by height in metres squared. Since each respondent had either two or three measures of height and weight each, we used the average of all recorded measures. There was no pre-population of this variable in wave 2.

- **Age** - This was measured in integer years. The variable triggered a data confirmation question for interviewers if the respondent had aged by less than 1 year or by more than 2 years between wave 1 and wave 2. Interviewers had access to the wave 1 roster, hence even fabricated surveys would likely have appropriate data in wave 2.

- **Years of education** - This variable is bounded between 0 and 15. The wave 1 information on education was also given to interviewers. Moreover, if the education levels had increased by more than 2 years, or had decreased between wave 1 and wave 2, the soft-

[29]We collapse the four labour market states into a binary employed variable for the regressions. We did this as it made more sense theoretically and it made the discussion of the regression results simpler. We also performed the estimations with the labour market states disaggregated and the overall findings do not change substantially (not reported).

ware would ask for confirmation from the interviewers. Thus, we expect to have only a small amount of measurement error on this variable in the fabricated data.

- **Male** - This is an indicator variable that captures the sex of the respondent. It was pre-populated based on the wave 1 dataset.

The labour market status variables are comprised of four mutually exclusive indicator variables,[30] which represent the labour market state of respondents. These are all derived from the labour market section of the survey. These variables were not pre-populated. They are:

- **Employed** - This is an indicator variable that takes on a value of one if the respondent had any form of employment at the time of the interview.

- **Unemployed (searching)** - This is an indicator variable that takes on a value of one if the respondent was not employed but was actively looking for work in the month prior to the interview.

- **Unemployed (discouraged)** - This is an indicator variable that takes on a value of one if the respondent was not employed and was not actively looking for work in the month prior to the interview, but stated that he/she would like to have a job. The difference between the searching and non-searching unemployed conforms to the standard ILO definitions for these categories (International Labour Office, 2011).

- **Not economically active** - This is an indicator variable that takes on a value of one if the respondent was not employed, was not actively looking for work in the month prior to the interview and stated that they would not accept a job offer.

There are a few additional data issues that require elaboration. First, for our entire analysis, we restrict our estimation sample to include only the adult African sub-population aged 18 to 65.[31] Second, we restrict the sample to include BMI values of less than 50, as we were concerned that some of the extremely high BMI values were due to the scales being inadvertently

[30] There are some cases where the questions used to derive a person's labour market status were not answered. In these cases, we cannot define their status. Otherwise, the four variables would be mutually exclusive and exhaustive.

[31] This is because we have small sample sizes for the other race groups, especially when using the balanced panel members from wave 1 and wave 2. Wittenberg (2013) applies a similar restriction to the NIDS data.

set to pounds instead of kilograms. In addition, we exclude any observation with any covariate missing from our samples, as they would not survive into our regression analyses. Third, we do not make use of either the sampling weights or attrition-corrected weights in any of the subsequent analysis. Our objective is not to replicate population level analyses, but merely to compare the differences between estimates obtained from the Dirty and the Clean dataset. Moreover, we would have had to recalculate all of the post-stratification weights, as the datasets that we use do not represent the full sample due to the time at which we completed our audit. Fourth, in our regressions we re-weight the subsequently corrected data by the inverse of the ratio of the number of corrected fakes to the number of fakes. We do this because we want the weighted fraction of data from the 'Always Correct' data to be the same in the Dirty and Clean datasets. The implicit assumption here is that the group of corrected fakes are representative of the group of fakes that we were unable to subsequently re-interview.[32]

The sample sizes, and how they are affected by our restrictions, are displayed in Table 2.7 below. We observe that the BMI cutoff of 50 is not too onerous. We lose 106 and 84 observations from the Dirty and Clean datasets respectively. This represents less than two per cent of either sample, and a substantial fraction of these are observations from the 'Always Correct' subset of the data. Our final sample sizes for the OLS and First Differences analysis are 6 768 and 5 388 observations for the Dirty dataset, and 6 576 and 5 263 for the Clean dataset, respectively. The sample sizes for the First Differences regressions are substantially smaller for two reasons. First, new household members would not have been interviewed in wave 1. Second, any missing data in any covariate in wave 1 would have resulted in that observation being dropped from the First Differences sample as well.

[32]We test the assumption that the corrected fakes are representative of all fakes by testing for the equality of means between the corrected fakes and the uncorrected fakes, from the corresponding wave 1 data, for the variables used in the regression section of the chapter. We find that we are not able to reject the hypothesis of equal means for any of these variables at the 10% level of significance.

Table 2.7: Sample sizes

Number	Dirty	Clean
All	6 874	6 660
BMI<50	6 768	6 576
Cross-sectional OLS	6 768	6 576
First differenced	5 388	5 263

Source: Own calculations using pre-public release NIDS Wave 2 data, 2010.
Sample restricted to African adults aged 18 to 65 in wave 2.

2.4.2 Summary statistics

The means of the variables in each of the sub-datasets, as well as the Clean and Dirty datasets, are provided in Table 2.8 below. Note that the mean of a variable in the Dirty dataset will be a weighted average of the corresponding means in the Fake dataset and the Always Correct dataset, with the weight being determined by the proportion of the data in the Dirty dataset that originates from the Fake and Always Correct datasets, respectively. Similarly, the mean in the Clean dataset will be a function of the means in the Corrected Fake and Always Correct datasets. Any differences in the means between the Clean and Dirty datasets must therefore reflect differences in the means between the Fake and Corrected Fake datasets, combined with the differences in their respective sample sizes.

Table 2.8: Means of variables used in analysis

	Fakes	Corrected Fakes	Always Correct	Dirty Dataset	Clean Dataset
BMI	26.73	25.76	26.94	26.92	26.89
Age	35.45	37.08	36.16	36.11	36.20
Education (years)	6.74	7.47	8.12	8.03	8.09
Employed	19.82%	28.40%	33.99%	33.05%	33.77%
Unemployed (searching)	5.12%	15.56%	10.98%	10.59%	11.16%
Unemployed (discouraged)	0.69%	3.50%	5.93%	5.59%	5.84%
Not economically active	73.27%	52.53%	48.60%	50.24%	48.75%
Male	40.98%	39.69%	40.26%	40.31%	40.24%
Number	**449**	**257**	**6 319**	**6 768**	**6 576**

Source: Own calculations using pre-public release and public release NIDS Wave 2 data, 2010.
Samples are restricted to Africans aged 18 to 65 in wave 2, with BMI values less than 50.
The number of fakes do not equal the number of corrected fakes because not all faked respondents were successfully re-interviewed. The means in the Dirty and Clean dataset do not precisely correspond to the weighted means obtained from the first three columns due to rounding effects.

For the BMI, age, male and years of education variables that we use, the difference in means between the Fake and Corrected Fake datasets is relatively small. Thus, in the aggregated Dirty and Clean datasets, the difference in means for these variables is very small, at less than 0.15 units in each case. For some of the other variables, such as Unemployed (discouraged), the difference in means between the Fakes and Corrected Fakes is somewhat larger, at 2.81 percentage points, but the aggregate difference in means for these variables remains relatively small. This is because the relative weightings contributed by Fakes and Corrected Fakes to the means in the Dirty and Clean datasets are also small.

In contrast, for the Employed, Unemployed (Searching) and NEA variables, the difference in means between the Fakes and Corrected Fakes is substantial. Even these differences, however, get substantially moderated when we calculate the means of the Dirty and the Clean datasets. For example, let us consider the percentage that are employed in the Fakes and Corrected Fakes datasets. The difference in the means is large, at 8.58 percentage points. Nonetheless, the weight that these contribute to the Dirty and Clean datasets is relatively small, at 6.6 and 3.9 percent. Thus the aggregate difference in mean percentage employed between the Dirty and Clean datasets is only 0.72 percentage points. This is substantially smaller than the corresponding difference in means between the Fakes and Corrected Fakes dataset, and depending on one's interest, may or may not be considered to be 'substantial'. Overall then, we confirm the finding by Schnell (1991) that a univariate statistic such as the mean is generally robust to the presence of a small amount of fake data.

2.4.3 Transition matrices

We next consider labour market mobility by calculating transition matrices across labour market states. Each row in the transition matrix in Table 2.9 contains the distribution of labour market states observed for respondents in wave 2, conditional on their wave 1 state. For example, of the 2 348 people who were not economically active (NEA) in wave 1 in the Dirty dataset, 73% were NEA in wave 2, while 12.27% had found employment.

The differences between the Dirty and Clean datasets are contained in the third matrix in Table 2.9, below. A clear pattern emerges, even though the magnitudes are not too large in absolute value. The presence of the faked data would cause us to systematically overstate the likelihood of transitioning into the NEA state, and underestimate the probability of transitioning or remaining in any of the other states, regardless of the initial wave 1 state. This reconciles well with the cross-sectional differences in means from Table 2.8, where the difference in the mean percentage that were NEA between the Fakes and Corrected Fakes was more than twenty percentage points, and the corresponding difference between the Clean and Dirty means was about 1.5 percentage points. It also seems consistent with the time-saving hypothesis that we postulated would accompany cheating behaviour.

Nonetheless, the magnitudes of the differences are generally below 1 percentage point, and only one entry is above 2 percentage points. Whether one considers these differences to be material or not will, once again, depend on one's perspective. If one were interested in population-level dynamics, or in long run forecasting of retirement behaviour, then they could well be material. On the other hand, if one were simply interested in the conditional probabilities of transitioning across labour market states, then the Dirty and Clean data would not have yielded a very different understanding of the aggregate levels of churning in the South African labour market.

Table 2.9: Transition matrices across labour market states (%)

		Wave 2 State (Dirty Dataset)					
		NEA	Unemployed (Discouraged)	Unemployed (Strict)	Employed	Total	N
W1 State	NEA	73.00	6.05	8.69	12.27	100	2 348
	Unemp. D.	53.13	10.34	12.26	24.28	100	416
	Unemp. S.	43.32	6.62	22.42	27.64	100	861
	Employed	30.16	4.31	7.48	58.05	100	2 205
		Wave 2 State (Clean Dataset)					
		NEA	Unemployed (Discouraged)	Unemployed (Strict)	Employed	Total	N
W1 State	NEA	71.47	6.32	9.58	12.63	100	2 296
	Unemp. D.	52.06	10.65	12.59	24.7	100	413
	Unemp. S.	41.04	6.96	23.47	28.54	100	848
	Employed	29.47	4.38	7.53	58.62	100	2 192

		Difference Dirty-Clean (%) Wave 2 State			
		NEA	Unemployed (Discouraged)	Unemployed (Strict)	Employed
W1 State	NEA	1.53	-0.27	-0.89	-0.36
	Unemp. D.	1.07	-0.31	-0.33	-0.42
	Unemp. S.	2.28	-0.34	-1.05	-0.90
	Employed	0.69	-0.07	-0.05	-0.57

Source: Own calculations using pre-public release and public release NIDS Wave 2 data, 2010.

2.4.4 Regression results

Our final set of analyses involves estimating the regression coefficient of employment on BMI. We first present the cross-sectional regression results using wave 2 data, and compare the coefficients from the Dirty and Clean datasets. One weakness of this approach, if we think that BMI is a proxy for health, is that we are likely to have a selection problem since healthier people will probably be more likely to find employment. A natural extension would be to estimate a fixed effects model of employment on BMI. We thus also present the First Differences (FD) form of the regression.

We do not have strong priors regarding the differences between the Dirty and Clean datasets based on econometric theory. The measurement error in the employment dummy cannot be classical measurement error as the variable is a binary variable (Aigner, 1973). Moreover, in the FD model, the error has a particular distribution that is not symmetric. Nonetheless, we do expect that we have a measurement error problem due to the misclassification of the employment dummy, and that this will cause an endogeneity problem because this misclassification enters the error term in the estimating equation. We also know that the FD estimator is more sensitive to measurement error than the OLS estimator, so we might expect that the presence of fabricated data will have a stronger effect on the FD coefficients than the OLS coefficients (Hausman, 2001).

In Table 2.10, we present the regression outputs from estimating the OLS model on the Dirty and Clean data. Our dependent variable is BMI and our regressors are age, gender, education and employment status. The overall finding is that the regression results look very similar. The R-squared values differ by 0.01 units, and none of the coefficients are statistically significantly different at any reasonable significance level. The education variable, where we do have something that resembles classical measurement error within a limited range, is indeed slightly smaller in the regression using the Dirty dataset, but the difference is only 0.03 BMI units, which is quite negligible. The coefficients on the employed dummy are both positive and significant. The coefficient is slightly larger when using the Dirty dataset (0.76 versus 0.70), but again they only differ by 0.06 BMI units.

Table 2.10: Cross-sectional and first-differenced regressions

	Cross sectional		First differenced	
Variables	**Dirty W2 BMI**	**Clean W2 BMI**	**Dirty Δ BMI**	**Clean Δ BMI**
Age	0.14***	0.14***	0.00	0.00*
	(0.01)	(0.01)	(0.00)	(0.00)
Male	-4.52***	-4.60***		
	(0.15)	(0.15)		
Education	0.13***	0.16***	0.08**	0.02
	(0.02)	(0.02)	(0.04)	(0.04)
Employed	0.76***	0.70***	0.31**	0.18
	(0.16)	(0.16)	(0.15)	(0.14)
Constant	22.50***	22.00***	1.03***	1.00***
	(0.37)	(0.38)	(0.08)	(0.07)
Observations	6,768	6,576	5,388	5,263
R-squared	0.19	0.20	0.00	0.00

Standard errors in parentheses
*** $p<0.01$, ** $p<0.05$, * $p<0.1$

Source: Own calculations using pre-public release and public release NIDS Wave 2 data, 2010. Samples are restricted to Africans aged 18 to 65 in wave 2, with BMI values less than 50. Columns 2 and 3 present results from cross-section OLS estimation. Columns 4 and 5 present results first-differenced regressions, and the regressors should be read as differences, rather than levels.

The similarities are not surprising given what was observed in Table 2.8. In the cross-sectional datasets the differences in the means of the relevant variables were all very small, and most of the data in both the Dirty and Clean datasets are obtained from the Always Correct dataset.

Our final set of results are obtained from the FD regressions and are presented in the last two columns of Table 2.10. Note that the male dummy gets dropped as it is a time invariant variable. Our findings from this component of our analyses are a bit more nuanced than those from our earlier analyses.

When we compare the differences in the FD results between the Dirty and Clean datasets, we notice that the Dirty coefficients on education and employment are larger than those obtained from the Clean dataset, and they are statistically significant whereas those obtained from the Clean dataset are not statistically significant. On the other hand, the differences in magnitude are 0.06 and 0.13 BMI units for the education and employed variables, which are not particularly large. Moreover, the differences in the coefficients are not statistically significant. From this perspective, the fabricated data does affect our estimates, but not in a meaningful way.

Alternatively, when we compare the OLS and FD results within each dataset, we observe that the coefficients from both the Dirty and Clean datasets are reduced quite substantially. For example, in the Dirty regressions, the coefficient on education is 0.13 in the OLS regression but decreases to 0.08 in the FD regression. The decrease obtained in the Clean dataset is from 0.16 to 0.02, and also results in a change in the statistical significance of the coefficient.

A similar comparison between the OLS and FD coefficients focusing on the employed dummy yields larger decreases in the coefficients (in absolute value) for both the Dirty and the Clean datasets. If our only dataset had been the Dirty dataset, we would have concluded that using a longitudinal estimator results in a decrease of our estimated coefficient from 0.76 to 0.31, that is a decrease of 0.45 BMI units, although both coefficients are statistically significant at the 5% level. In contrast, if we had performed the identical exercise using only the Clean dataset, we would have concluded that using a longitudinal estimator results in a decrease of our estimated coefficient from 0.70 to 0.18, that is, a decrease of 0.52 BMI units. Moreover, we would observe that the FD coefficient, unlike the OLS coefficient, is not statistically significantly different from zero.

One question that comes to mind is why the coefficients in the regression output are larger for the Dirty data than for the Clean data. If there is classical measurement error in the X variables, then one expects the point estimates to be biased downwards towards zero. However, in our study, the opposite is true. There are a number of reasons why this may be the case.

First, the results of regressions on the Dirty dataset presented in Table 2.10 are not subject to the usual assumptions about classical measurement error. For example, the education variable approximates a continuous variable, but given the computer check when recording this variable, it is not clear what the realized measurement error will look like. Second, the Dirty dataset includes measurement error of the dependent variables as well. With classical measurement error affecting a dependent variable, we would expect larger standard errors corresponding to the coefficient estimates, but our point estimates should be unaffected in expectation. However, the possible non-classical measurement error on the Y variable in our study means that this expectation is no longer valid. In addition, the setting of our study is different to the majority of the literature on measurement error. Instead of presenting a model in which we consider the error of an X variable or a Y variable in isolation, the Dirty dataset potentially has non-classical measurement error in both the dependent and independent variables simultaneously. The possible correlation between errors associated with these variables means that the standard framework is not applicable to this study. Extending from the cross section to the first differenced regressions, it is also unclear what one should expect if one begins with non-classical measurement error in the X variables and then estimates a longitudinal model using these variables.

2.4.5 Discussion

We are aware that we are probably not getting the true 'causal' estimate of labour market status and change in labour market status on BMI, but our focus is primarily on measuring the difference between the estimates obtained by using the fabricated data instead of the subsequently corrected data. This is the main contribution of this section of the chapter. To our knowledge, all previous research on this topic has amounted to comparing an *ex ante* dataset containing fabricated data to an *ex post* dataset where the fraudulent data has been deleted, but not replaced. In our case, where households with fabricated data were re-integrated into the NIDS dataset, we are in a unique position in that we observe both the fabricated data as well as the subsequently corrected data.

The overarching question that we set out to answer was whether the fabricated data would have affected our estimates in a meaningful way. Our findings suggest that the answer to this question depends on the estimator being considered and the purposes for which the analysis is being conducted. The general picture that emerges, which is consistent with econometric theory, is that the cross-sectional estimates of means and OLS regressions are not substantially affected by the presence of a relatively small amount of fabricated data. At the same time, the identical amount of fabrication does affect the longitudinal estimators. For the FD regressions, the difference in the estimates would have led us to reach quite different overall conclusions.

It is useful to try to perform a cost-benefit analysis of our data quality investigation. Challenges arise in terms of attributing financial costs to the various activities that were undertaken, as well as on placing a financial value on the better quality data. Nonetheless, even the most conservative calculations suggest that the investigation was worthwhile in that the value of the benefits was at least 24 times the aggregate costs associated with the related activities. For example, the overall budget for the entire wave 2 operation was approximately R34 million.[33] If the deliverable of the entire project is a dataset, and we use the 7.3% fabrication rate that we identified, and we assume that fabricated data has no value, then the cost of the fabrication would have been about R2.48 million rand. In contrast, the costs associated with the investigation were primarily driven by the time invested by various members of project staff. In total, we estimate the salary costs to have been not greater than R100 000. This results in a 'benefits-to-cost' ratio of greater than 24, and we believe that this is a conservative estimate.[34] As such, undertaking active steps to quickly identify cheating interviewers is easily justifiable.[35]

Better quality data is always desirable. Thus, other things being equal, we should always get the best quality data that we can. Unfortunately, running a survey is a costly and complex task, and the costs of auditing and monitoring interviewers competes with several other tasks for resources in a finite budget. Thus, the 'other things being equal' assumption is not very realistic. Nonetheless, given the overall costs and effort invested in running a large survey, the marginal cost of performing some generic data quality checks for fabricated data seems to be highly warranted, especially in an environment where we now have evidence of fabrication in several studies. In our study, the estimators that were most affected were the longitudinal estimators. At the same time, the marginal cost of detecting fabricated data can be substantially lowered when one has longitudinal data, as one can then look for inter-temporal data anomalies in addition to cross-sectional data anomalies. This makes an even stronger case for the argument that adequate resources be allocated for identifying data fabrication in longitudinal studies.

---

[33] At the existing 2010/2011 exchange rate, this was approximately 4.25 million US dollars.

[34] The estimate is conservative for several reasons. First, we have been generous in accounting for the time spent by the staff on these specific activities, as most staff were working on several other tasks at the time. Second, the proportion of fabricated data would potentially have increased with time in the absence of our investigation, as the cheating interviewers were completing interviews at a relatively fast rate. Third, assuming that fabricated data has no value is itself conservative. The point of the survey is to obtain data that will assist in evidence based policy making, and thus fabricated data, insofar as it leads to false research findings, has the potential to have a strictly negative and potentially large value.

[35] The benefits-to-cost ratio calculated in this section is from the perspective of the survey organisation (NIDS) and not from the perspective of the company implementing fieldwork.

2.5 Conclusion

In this chapter, we argued that the incidence of interviewer cheating is widespread. We documented cheating and potential cheating in five substantial South African surveys. Of the various methods that we considered to detect fraudulent data, two were more useful in our context than the others. These were Benford's law and the identification of anomalies in the anthropometric data of respondents.

Looking forward, there may be ways to improve on our process for identifying fabrication. First, survey companies that are using computers with built-in GPS devices to fill in questionnaires can use the software to capture the time and place that a survey was conducted. This can be done without the knowledge of the interviewer, which will aid in detection. Using a GPS software will allow much better monitoring of interviewers' whereabouts while they are in the field. In addition, interviewers that fabricate data are likely to complete entering the data much faster than an actual survey would take to complete. These two pieces of information alone would greatly improve the data quality auditing process.

Second, wireless networks and cellular technologies are now widespread even in developing countries. It is thus not unrealistic to expect to get data with just a day's lag. Previously, while using paper questionnaires, it would take months to obtain data in an electronic form. With the real-time uploading of the data, one can now check on each interviewer much earlier in the process, and constantly monitor each interviewer's performance. This enables survey organisations to fire cheating interviewers, as well as compel the fieldwork company to redo the interview.

The incentive structure facing interviewers is also an important part of reducing the probability of data fabrication, even before going to field. It is easy to see how a system of paying interviewers immediately for completed questionnaires could lead to higher rates of fabrication. These perverse outcomes could be mitigated by delaying payment to interviewers until a certain proportion of each batch of completed questionnaires has been verified. This incentive structure is also reflected in potential principal-agent problems between the organisation commissioning data collection (the research organisation) and the organisation responsible for collecting the data (the fieldwork company). One possible way of mitigating this is to assign funds in the fieldwork budget that are specifically dedicated to verification checks. Another is to make part of the payment to the fieldwork company conditional on the successful implementation of fraud minimisation and data verification exercises. Given that surveys such as NIDS are longitudinal and require data collection over multiple years, the potential for repeated interactions between the research organisation and the fieldwork company also serves to align incentives.

One important part of obtaining high quality data relates to interviewer selection and training. In NIDS, fieldwork was outsourced to a survey company under the condition that all interviewers had at least completed secondary school. In addition, all interviewers were required to attend an intensive week-long training course, during which they received repeated assessments and feedback. Fieldworkers who did not meet the required standards were not allowed to proceed with actual interviews. Despite this, the probability of fabrication cannot realistically be assumed to be zero. Thus, *ex post* data quality checks remain an integral process with which we can verify the data quality. This should be thought of as complementary to the other efforts at maintaining data quality that occur prior to fieldwork.

Other possibilities might be to use built in cameras to take photographs of survey respondents, real-time callbacks to ensure that the interview did in fact take place, and to strategically not pre-populate certain variables in longitudinal studies so that sizable deviations from time-invariant or slow-moving variables are flagged immediately. In summation, it seems that there are several relatively low cost ways in which survey organizations can use modern technology to minimise both the likelihood of interviewer cheating, as well as the impact of such cheating on the overall quality of the data, and explicitly performing such quality control activities is easily justifiable.

Appendix

2.A Time Use Study

Table 2.A.1: Time Use Survey selection grid

Persons 10 years +	HH1	HH2	HH3	HH4	HH5	HH6	HH7	HH8	HH9	HH10
1	1	1	1	1	1	1	1	1	1	1
2	1 2	1 2	1 2	1 2	1 2	1 2	1 2	1 2	1 2	1 2
3	1 2	1 3	2 3	2 3	1 3	2 3	1 2	2 3	1 3	2 3
4	2 4	1 3	1 3	2 4	1 3	2 4	1 3	2 4	1 3	2 4
5	3 5	1 4	1 3	2 4	1 5	2 4	2 4	4 5	1 2	2 4
6	5 6	4 6	1 2	1 2	1 5	4 6	1 5	3 5	4 6	1 3
7	2 6	4 6	2 5	5 7	2 4	4 7	5 7	1 4	2 6	1 4
8	1 5	1 3	6 8	2 5	1 4	5 6	2 3	5 7	6 8	2 8
9	4 9	1 3	4 9	1 5	2 7	2 9	2 3	4 5	7 8	2 6
10	3 9	1 6	2 3	4 9	1 3	8 10	5 6	3 7	2 5	8 9

Source: Time Use Survey Fieldworker Manual.

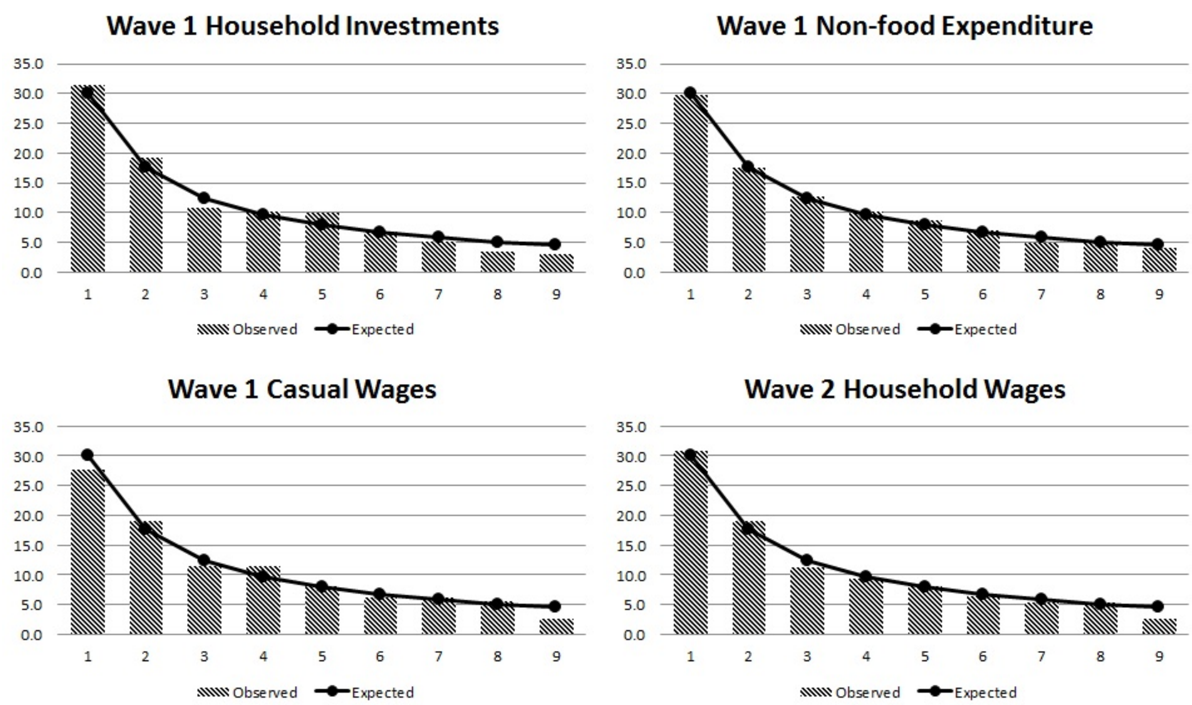
Table 2.A.2: Chi-squared tests for difference in distributions

Number Eligible	Number of HHs		Person Number 1	2	3	4	5	6	Total	Chi-sq.
3	1 045	Expected %	50	70	80					
		Expected #	523	732	836				2 090	$\chi^2(2)$
		Actual #	845	844	401				2 090	CV: 5.99
		Difference	323	113	-435				0	Test stat: 442.7
4	1 104	Expected %	50	50	50	50				
		Expected #	552	552	552	552			2 208	$\chi^2(3)$
		Actual #	738	731	530	209			2 208	CV: 7.91
		Difference	186	179	-22	-343			0	Test stat: 334.7
5	901	Expected %	40	50	20	60	30			
		Expected #	360	451	180	541	270		1 802	$\chi^2(4)$
		Actual #	529	434	476	265	98		1 802	CV: 9.45
		Difference	169	-17	296	-276	-172		0	Test stat: 815.4
6	590	Expected %	50	20	20	30	40	40		
		Expected #	295	118	118	177	236	236	1 180	$\chi^2(5)$
		Actual #	293	214	263	193	140	77	1 180	CV: 11.1
		Difference	-2	96	145	16	-96	-159	0	Test stat: 403.9

Expected numbers rounded to closest integer.

2.B Leading digits of other monetary variables

Figure 2.B.1: Leading digit distributions for other monetary variables



Source: Own calculations using NIDS Wave 1 2008 and Wave 2 2010/2011.

2.C Full fieldworker Chi-squared table

Table 2.C.1: Interviewers by Chi-squared ranking

Ranking	Interviewer	Chi-squared (no. of interviews)
1	**A**	**39.7 (80)**
2	**B**	**31.2 (49)**
3	OK	28.0 (66)
4	**C**	**27.3 (74)**
5	**D**	**27.1 (100)**
6	**E**	**24.7 (42)**
7	OK	21.7 (64)
8	**F**	**21.4 (73)**
9	OK	21.2 (44)
10	OK	19.8 (67)
11	OK	19.7 (43)
12	**G**	**18.7 (47)**
13	OK	18.0 (58)
14	OK	18.0 (73)
15	OK	17.0 (49)
16	OK	16.0 (52)
17	OK	15.9 (53)
18	OK	15.8 (53)
19	OK	14.9 (81)
20	OK	13.6 (41)
21	**H**	**13.2 (51)**
22	**I**	**12.3 (53)**
23	OK	11.8 (48)
24	OK	11.5 (56)
25	OK	10.9 (64)
26	OK	10.7 (48)
27	OK	10.5 (43)
28	OK	10.3 (44)
29	OK	8.4 (51)
30	OK	8.1 (41)
31	OK	7.2 (62)
32	OK	6.7 (57)
33	OK	6.4 (82)
34	OK	6.3 (41)
35	OK	6.1 (49)
36	OK	5.8 (66)
37	OK	5.7 (47)
38	OK	4.5 (71)
39	OK	4.5 (58)

Source: Own calculations using pre-public release NIDS Wave 2 data, 2010.
Interviewers who were found to have fabricated some or all of their data are emphasized in bold.

2.D NIDS verification questionnaire

Figure 2.D.1: NIDS verification questionnaire

INDIVIDUAL QUESTIONNAIRE

To be used only if MAIN respondent answered NO, to either Q11a or b (Was not present or unsure of question details)

HHID Number:	**PID Number:**

		Yes/Correct (✎✓)	**No/Incorrect** (✎✓)
	INDIVIDUAL QUESTIONS		
1.	Were you interviewed in person by an interviewer for the NIDS survey in 2010?	☐ (Ask Q2a)	☐ (Ask Q7)
2a.	**READ OUT:** We would like to confirm some of the questions, to see if we have the correct information on our system. Did the interviewers ask you personally you were born and where you might have lived previously?	☐ (Ask Q2b)	☐ (Ask Q2b)
2b.	Where were you born? **Interviewer Note:** Open UMPC to birth history and confirm where the respondent was born.	☐ (Ask Q2a)	☐ (Ask Q2a)
3a.	Did they ask you detailed questions about your mother and father, such as when they were born and what their last work was?	☐ (Ask Q3b)	☐ (Ask Q3b)
3b.	When was your mother born? **Interviewer Note:** Confirm the details on the UMPC Is the answer recorded the same as the answer given by the respondent?	☐ (Ask Q4a)	☐ (Ask Q4a)
4a.	Did they ask you questions about your education, such as the grades you completed and any other studies you may have done after school?	☐ (Ask Q4b)	☐ (Ask Q4b)
4b.	What is the last school you attended? **Interviewer Note:** Confirm the details on the UMPC Is the answer recorded the same as the answer given by the respondent?	☐ (Ask Q5)	☐ (Ask Q5)
5.	What your employment status at the time you were interviewed? Were you employed, unemployed or self employed? **Interviewer Note:** Confirm the details on the UMPC Is the answer recorded the same as the answer given by the respondent?	☐ (Ask Q6)	☐ (Ask Q6)
6.	Were your measurements taken, i.e. your height, weight, waist and blood pressure?	☐	☐ (Ask Q7)
7.	**Interviewer Note: ONLY** If No/Incorrect recorded in any of the questions above, **READ OUT:** We seem to have some missing information. Would it be all right if we completed these sections now of could we come back at another more convenient time to you and other household members?	☐ (Proceed to individual interview/sections on UMPC dependant on information omitted/incorrect)	☐ (Ask Q7)

3 The dynamics of poverty in South Africa

3.1 Introduction

There is general consensus that the extent of money-metric poverty in South Africa has declined over the last decade and a half. Various cross-sectional studies of poverty using household survey data have chronicled a decline in the poverty headcount that is largely attributable to the role of state support of household incomes (Leibbrandt et al., 2010; Bhorat et al., 2012; Leibbrandt et al., 2012; Leibbrandt and Levinsohn, 2016). However, far less is known about contemporary poverty dynamics in the country.

In the current South African policy milieu there is a rising emphasis on understanding how and why people enter and exit poverty. The aim of this chapter is to investigate the dynamics of poverty in South Africa using the first four waves of the National Income Dynamics Study (NIDS). The focus is on absolute, rather than relative, poverty transitions. In this chapter absolute poverty refers to using absolute poverty lines when analysing poverty dynamics. That is, a line is chosen that is agnostic both to how many people are above or below it, and to its position in the distribution of income. This line is held constant in real terms over the different years for which data are available. A relative poverty line (for example some fraction of mean or median income) is not used. The use of an absolute rather than a relative poverty line is preferred as it is based on a cost-of-basic-needs approach, which attempts to quantify the minimum amount of household per capita income required to cover basic food and non-food costs. This is thought to be more informative than a relative line such as 50% of the median, given how unequal South African society is, and given the large proportion of South Africans who fall below any reasonably defined absolute poverty line. Furthermore, the focus on absolute rather than relative poverty lines is in keeping with much of the preceding poverty literature in South Africa, and allows for a more natural comparison with previous findings. One of the key features of NIDS is the ability to model these dynamics of poverty over time. We are less interested in describing cross-sectional poverty and more interested in understanding the extent of movements into and out of poverty, who is making these transitions and the reasons for these changes.

This chapter has two distinct sections. In the first we present a wide-ranging descriptive analysis of poverty in South Africa using the balanced panel sample of respondents from the first four waves of NIDS. In doing so we uncover how much absolute mobility (both upwards and downwards) was experienced by balanced sample members between 2008 and 2014/2015. We also document the poverty-time interaction by breaking down poverty into chronic and transitory components. In the second section we move beyond the simple enumeration of poverty by shifting our focus to an econometric analysis of welfare dynamics using the full sample available to us over the four waves of NIDS. In doing so, the focus changes from an analysis of four wave poverty transitions using the balanced panel to an analysis of transitions using a pooled dataset of all respondents. Our modelling strategy allows us to model poverty dynamics while controlling for initial conditions and non-random attrition. It also allows us to separate genuine state dependence from aggregate state dependence, and this has potentially important policy implications.

This chapter contributes to the existing South African literature by being the first to use nationally representative data to study changes in money-metric welfare over this extended time period. It also takes seriously the questions of how to model these dynamics. This means being cognizant of two phenomena that need to enter our estimation. First, we ask how important is state dependence - whether or not an individual is initially poor or non-poor - in determining poverty dynamics. Second, we ask how important is selective attrition in our panel sample in determining poverty dynamics. It is also the first South African study that attempts to separate genuine state dependence from aggregate state dependence - that is, it measures the relative importance of initial poverty status versus individual-level unobserved heterogeneity in driving dynamics.

Section 3.2 of this chapter follows Finn and Leibbrandt (2016) by briefly outlining the South African literature on poverty dynamics,[1] and section 3.3 discusses the data and weights used in the first part of the analysis. Section 3.4 develops a number of univariate and multivariate measures of poverty transitions, with inter-wave poverty entry and exit being treated separately. Section 3.5 elicits the relative contributions of trigger events that are associated with poverty transitions. The chapter shifts focus to a Markovian model of poverty transitions in section 3.6,

[1]The working paper cited undertakes an earlier analysis of some of the transitions contained in the first section of this chapter, but with only the first three waves of NIDS. This chapter extends that analysis by considering data and transitions over a longer period of time, and by explicitly modelling poverty transitions over the 2008 to 2015 period in a Markovian framework.

while section 3.7 provides some concluding remarks.

3.2 The South African literature on poverty dynamics

Although there have been many studies of cross-sectional poverty in South Africa since the end of apartheid (see Finn, Leibbrandt and Ranchhod (2014) for a short review), there is a relative paucity of literature using panel data to analyse transitions. The best known study of poverty dynamics in post-apartheid South Africa is Carter and May (2001). The authors use the first two waves (1993 and 1998) of the KwaZulu-Natal Income Dynamics Study (KIDS) to decompose poverty transitions into what they term 'structural' and 'stochastic' components, using a sample of approximately 1 200 African households in the KwaZulu-Natal province. The authors find a significant increase in poverty rates in African households in the province, and also find that the economic processes driving poverty dynamics served to increase inequality. That is to say, upward economic mobility was stronger for those at the top of the income distribution than it was for those at the bottom. The authors find that approximately one fifth of the sample was poor in both 1993 and 1998, with a further 35% transitorily poor (that is, poor in at least one wave).

Woolard and Klasen (2005) also use the first two waves of KIDS to model the determinants of mobility and poverty transitions for just over 1 000 African households in KwaZulu-Natal. The authors identify the main event associated with a transition into or out of poverty in a univariate sense. These events are themselves split into demographic (household composition) changes and income changes. It is found that about one quarter of transitions into and out of poverty are due to demographic effects. The most important income effect for transitioning into poverty is the household head losing a job, while the most important income event for transitioning out of poverty is another household member finding employment. The importance of demographic effects is confirmed in a multivariate regression analysis, though the sample sizes are quite small with 129 households entering poverty and 223 households exiting poverty over the two waves.

Agüero et al. (2007) add the third (2004) wave of KIDS to the study of dynamics. Parts of the paper are a natural update to Carter and May (2001), as the third wave is added as a new data point. The authors complement the income analysis by calculating poverty rates using expenditure data, though there are some serious misgivings about using the 1993 expenditure data (see Leibbrandt et al. (2010)). The study finds that access to basic household services improved significantly between 1993 and 2004, and this improvement is in contrast to the backward steps taken on the poverty front in the mid-1990s. Finally, the authors highlight the importance of government grants and, particularly, the child support grant, in shifting the bottom of the income distribution to the right, and find that the impact of grants as inequality reducers increased over time.

Finn et al. (2013) use the first two waves of NIDS to explore absolute and relative transitions over the 2008 to 2010/2011 period. They find that almost three quarters of those who were below the poverty line in 2008 were still below it in 2010/2011. This equates to approximately 34% of the total sample being poor in both waves for their poverty line. Poverty exits slightly outweighed poverty entries over the period, and this resulted in a small fall in the national poverty headcount ratio.

Finally, Finn and Leibbrandt (2013) expand on this previous study by adding a third wave to the analysis. They find that although the rate of exiting poverty was higher between waves 2 and 3 than between waves 1 and 2, a large percentage of the South African population was trapped in severe poverty (defined as living in a household with income per capita of less than half the poverty line) in all three waves. They also document that the reduction in non-money-metric (multidimensional) poverty was significantly larger than the concurrent reduction in income poverty over the period.

3.3 Data and summary statistics of the balanced panel

The data used in this part of the chapter come from the first four waves of NIDS, covering 2008-2014/15 (SALDRU, 2016*a*,*b*,*c*,*d*). The four waves of NIDS were collected in 2008, 2010/2011, 2012 and 2014/2015, respectively. NIDS is a nationally representative longitudinal dataset of individuals. Respondents are tracked over time, even if they change residence. In order to be considered a resident member of a household, an individual must usually reside in a dwelling unit for at least four nights a week, and must share food and resources from a common source with other household members. As the focus in this first part of the chapter is on describing poverty dynamics and transitions, the analysis is restricted to the balanced panel – those for whom we have complete interview data in all four waves.[2]

Although the focus of this chapter is on the use of micro data to understand poverty transitions, it should be noted that South Africa's macroeconomic environment between 2008 and 2015 was not conducive to the reduction of poverty. Table 3.1 presents GDP and GDP per capita numbers as well as their growth rates for the country covering the same period as the first four waves of NIDS. The effect of the recession which lasted from the final quarter of 2008 until the end of the second quarter of 2009 is clear, with GDP per capita shrinking by 2.7% in 2009. Even though the recession technically ended in 2009, growth rates thereafter were generally quite low and were barely above zero during the collection of the fourth wave of data. Trends in unemployment over the same period were equally concerning. Essers (2017) shows how unemployment increased as a result of the recession and remained high at least until the end of wave 3, and this was driven by reduced inflows of workers into jobs, rather than increase outflows of workers from jobs. The post-apartheid trend of the economy shedding low skilled jobs continued over the late 2000s and early 2010s, and the labour market failed to pull poor households into employment (Leibbrandt et al., 2016).

[2]The sample includes adults and children for whom we have interview information in all four waves. For children aged 0 to 14 a child questionnaire is administered to the mother or primary caregiver of the child, or to another household member who is knowledgeable about the child. An adult questionnaire is administered directly to respondents who are aged 15 and above.

Table 3.1: Macroeconomic trends in South Africa: 2008 to 2015

Year	GDP (ZAR million)	GDP growth (%)	GDP per capita	GDP per capita growth (%)
2008	2 708 600	3.2	54 322	1.9
2009	2 666 939	-1.5	52 838	-2.7
2010	2 748 008	3.0	53 823	1.9
2011	2 838 258	3.3	54 968	2.1
2012	2 901 076	2.2	55 543	1.0
2013	2 973 292	2.5	56 234	1.2
2014	3 023 826	1.7	56 469	0.4
2015	3 063 101	1.3	56 449	0.0

Source: Data from South African Reserve Bank (2017)

The timing of the collection of NIDS data raises two immediate points of concern. First is the relatively long intervals between waves. Given that the time between waves was generally more than a year (and, on average, more than two years), it must be the case that NIDS underestimates the prevalence of short spells of poverty. The average number of months between interviews for balanced panel members between wave 1 and wave2, wave 2 and wave 3, and wave 3 and wave 4 were 30.0, 21.4 and 30.6, respectively. The second concern is the unequal nature of the spacing between interviews over waves. For example, some of the balanced panel members had only 8 months in between being interviewed for wave 2 and wave 3. At the other extreme, other members of the balanced panel had 42 months in between being interviewed for wave 1 and wave 2.[3] It does not appear that differential intervals between waves are systematically related to different population subgroups, though, of course, the longer the time between interviews, the less likely we are to pick up short run dynamics.

Selective attrition over the successive waves of NIDS is something that has been a concern – see, for example, de Villiers et al. (2013) and Baigrie and Eyal (2013) who note the disproportionate loss of white respondents and relatively wealthy respondents (as separate from race) between the first and second waves of NIDS. The attrition rates for NIDS are relatively high compared to those in other countries with national longitudinal surveys (though these are almost all OECD countries). This is particularly true for the wave 1 to wave 2 interval. Chin-

[3]Figure 3.A.1 in the appendix presents the distributions of time in between interviews for balanced panel members for wave 1 to wave 2, wave 2 to wave 3, and wave 3 to wave 4.

hema et al. (2016) provides the overall attrition rates for wave 1 to wave 2, wave 2 to wave 3, and wave 3 to wave 4. These stand at 21.95%, 15.82%, and 13.75% respectively. These high attrition rates mean that constructing attrition-corrected weights for the balanced panel sample is an important undertaking. In order to adjust the balanced sample for the presence of selective attrition between waves 1 and 2, 2 and 3, and 3 and 4, we constructed a balanced panel weight. This was done by adjusting the original wave 1 post-stratified weight to account for unfolding attrition. For each successive wave a probit model was run with the dependent variable being a dummy indicating whether the individual attritted or not. Wave 1 to wave 2 balanced panel members then received a new weight which was the product of the original wave 1 weight and the inverse of the conditional probability of re-interview. The same process was applied to the wave 2 to wave 3, and the wave 3 to wave 4 periods. Given that the original wave 1 weight was multiplied by the inverse of the probability of re-interview, those belonging to groups that were more likely to attrit in between waves received a relatively higher weight. This means the attrition corrected weights are higher for, for example, white respondents, wealthier respondents, and the elderly, all of whom were relatively more likely than their counterparts to drop out of the sample, albeit for differing reasons. In some cases a high wave 1 calibrated weight was multiplied by a high attrition weight, resulting in an extremely high panel weight. In light of this, and in line with the NIDS methodology outlined in Chinhema et al. (2016), the panel weights were trimmed at the 1st and 99th percentiles.

There are 17 265 members of the balanced panel, and Table 3.2 presents some summary statistics for this subsample. 83% of our sample is African, with coloured and white proportions standing at about 8% and 7% respectively. The Indian part of the balanced panel is very small, with only 151 respondents being successfully interviewed in all four waves. For this reason, racial breakdowns including this group are generally avoided, because of the lack of power associated with such a small sample size.

As expected in a subsample that is ageing, the average level of educational attainment rose with each successive wave.[4] The large decrease in the proportion of the balanced panel with no education is because of the inclusion in wave 1 of children of all ages, many of whom started

[4]The education columns do not sum to 100% within each wave due to missing values for some respondents.

primary school over the course of the first four waves of NIDS. The share of the balanced panel with no schooling dropped from 20% in wave 1 to 6% in wave 4, and almost a quarter had obtained at least a matric by wave 4.

The evolution of the household size variable is interesting to observe. The share of the balanced panel living in single-person households rose by almost four percentage points between wave 1 and wave 4. This category and the next smallest (2 to 3) were the only two to grow between 2008 and 2015. The proportion of balanced panel members living in households with 4 to 6 people was 39% in wave 4, down from 44% in wave 1. The trend to smaller household sizes in the balanced panel is reflected in the cross-section as well. In the cross-section, average household size decreased from 3.53 in wave 1 to 3.20 in wave 4.

Turning to the three geo-types we see that the proportion of balanced panel members living in urban areas rose from 57% in wave 1 to 60% in wave 4, while the shares in traditional areas and farming areas decreased between 2008 and 2015. The provincial breakdown of balanced panel members was relatively stable over the period, with small decreases in the share living in the Eastern Cape and Limpopo, and a rise in the proportion living in Gauteng (not shown).

Table 3.2: Summary statistics of the balanced panel

	Wave 1	Wave 2	Wave 3	Wave 4
Race				
African		82.75%		
Coloured		8.23%		
Asian/Indian		2.34%		
White		6.67%		
Gender				
Male		47.04%		
Female		52.96%		
Age	26.47	28.89	30.70	33.21
Education				
None	20.33%	16.33%	12.71%	5.76%
Primary	32.73%	31.77%	30.75%	29.56%
Inc. Sec.	28.70%	31.88%	34.55%	38.57%
Matric	16.27%	17.92%	19.68%	22.25%
Tertiary	1.48%	1.71%	1.99%	2.39%
Household Size				
1	5.46%	6.17%	7.19%	9.25%
2-3	21.75%	19.59%	22.18%	23.94%
4-6	43.85%	41.71%	41.27%	39.04%
7-10	20.60%	23.85%	21.94%	20.61%
>10	8.33%	8.68%	7.43%	7.16%
Geo-type				
Traditional	37.55%	37.65%	36.82%	35.53%
Urban	57.23%	56.92%	58.44%	59.85%
Farming	5.22%	5.43%	4.75%	4.63%
N		**17 265**		

Source: Own calculations from the first four waves of NIDS. Attrition-corrected panel weights applied to members of the balanced panel.

Before presenting poverty transition matrices, it is worth spending some time looking at the composition of household income of poor and non-poor households in each of the four waves. Our welfare measure in this chapter is real monthly household income per capita.[5] We make use of the household income variable in the public-release dataset which was adjusted to remove imputed rent from owner-occupied housing in each wave. This was done because the imputed rent variable in each wave contained a high percentage of missing values, making it a very noisy component of income (even after single regression imputations were used to predict the missing

[5]We remain agnostic to the dynamics of non-monetary measures of well-being, which, by all accounts, have improved more rapidly than improvements in household income (Finn and Leibbrandt, 2016).

values). This follows the practice in many papers using the household income variable in NIDS (for example see Leibbrandt et al. (2010)). Disposable household income is defined as the sum over individuals in the same household of wages, remittance income, grants and income from investments. Wages include the net income received from primary jobs, secondary jobs, self-employment and casual work. Remittance income includes all monetary transfers received by the household from non-resident household members. Grant income includes the state old age pension, the child support grant, the disability grant, the care dependency grant, the foster child grant and the war veteran's grant. The sum of these components across all individuals in the household is then divided by household size in order to reach a measure of monthly household income per capita.

The choice of any equivalence scale when defining a welfare measure is going to involve some trade-offs. In this thesis, dividing total household income by the number of resident household members in order to reach the chosen welfare measure assumes that there are no economies of size within the household. This approach continues the precedent set in the analysis of poverty in South Africa (see, for example Woolard and Leibbrandt (2006)). Although the assumption of no economies of size is quite an extreme one, it is not clear that any of the alternatives offer a superior approach. For example, an equivalence scale such as the one used in Woolard and Klasen (2005) in which household income is divided by $(adults + 0.5 \times children)^{0.9}$. However, as noted in Budlender et al. (2015), defending the choice of these parameters is far from easy, and their use may raise more questions than answers. Woolard and Leibbrandt (2006) provides some evidence that poverty analysis in South Africa is generally robust to the choice of equivalence scale, although there is still some debate about the issue (see, for example, the analysis in Posel and Rogan (2016). Given the number of issues that would be raised by the choice of any equivalence scale, the chapter proceeds by appealing to precedent and to the simplest option, and simply divides total household income by the number of resident members.

We used Statistics South Africa's (StatsSA) headline CPI index to deflate the nominal income data to their real values. The base period is January 2015, as this was the modal month of interview for wave 4. All analysis that follows reports the income variables at their January

2015 price levels. The use of headline CPI assumes that price changes facing people in different parts of the country, and in different parts of the income distribution are the same. This is almost surely a simplification of reality, as in practice different households consume different bundles of goods, and therefore face different inflation rates. The use of plutocratic weights in the construction of CPI means that the fixed basket of goods underlying CPI is more representative of households with relatively higher expenditure. Leibbrandt et al. (2016) calculate percentile-specific inflation rates for South Africa and show that price changes between 2005 and 2010 were anti-poor. That is, the rate of inflation faced by poor households was higher than that faced by non-poor households, and also higher than the inflation reflected in headline CPI. However, it is far from clear that deflating prices for urban and rural households separately would lead to an improved measure of household income over time. As noted in Finn, Leibbrandt and Oosthuizen (2014), the rural price indices released by StatsSA are calculated on the basis of combining urban prices and rural expenditure weights. This creates a rather murky version of rural inflation, as it would only be accurate if the prices faced by rural households were exactly the same as those faced by urban households. For this reason we choose to use a single headline price index to deflate incomes over time. If one extends the findings in Leibbrandt et al. (2016) to the NIDS dataset, then choosing headline CPI rather than a percentile-specific CPI is likely to slightly underestimate the true prevalence of poverty in the country.

In Figure 3.B.1 in the appendix, we present eight bar charts - four for poor households in each wave and four for non-poor households in each wave. The y-axis represents the proportion of total household per capita income made up of each component of household income. These components are wages, government grants, remittances and investment income. Investment income comprises stocks, rentals, private pensions and retirement annuities.

A comparison of the share of wages in total household income for poor versus non-poor households shows that labour market income is substantially more prominent in the latter than in the former. The wage share in poor households for this period ranged from 46% to 52%, while in non-poor households it was stable at around 86%. The importance of income from government grants for poor households is clear in this figure, with the share of income coming from this source in the mid 40%s. As we show later, in Table 3.7, an increase in income from government grants was a very important trigger leading households to exit poverty between wave 1 and wave 4. Remittances play a more important part in the composition of income for poor households versus non-poor households - reaching a peak of 11% compared to 3% in wave 4. Finally, investment income makes up between 7% and 10% of household income for non-poor households in waves 1 to 4, compared to between 1% and 3% for poor households.

3.4 Descriptive poverty transitions

The NIDS wave 3 poverty transitions report (Finn and Leibbrandt, 2013) used a cost-of-basic-needs poverty line of R636 per capita per month (in August 2012 price levels) which itself was based on the line in Özler (2007). In this chapter, however, we use a poverty line that was derived by Budlender et al. (2015) of R1 283 in January 2015 rands.[6] This poverty line was calculated by first deriving a nutrition poverty line to reflect the minimum cost of a daily caloric intake of 2 100 kilocalories. This food poverty line was added to the average amount of non-food expenditure of households with food expenditure at the nutrition line in order to reach the amount of R1 283. In adjusting the original Budlender et al. (2015) line to its real January 2015 equivalent, we deflate the food and non-food components separately using CPI reports from StatsSA.

Transition matrices

In Table 3.3 we present poverty transition matrices for the balanced panel members. The four panels of the table show transitions from wave 1 to 2, 2 to 3, 3 to 4, and 1 to 4 respectively. Focusing on the wave 1 to wave 4 transition (shaded in grey) we see that of those balanced panel members who were poor in wave 1, almost three quarters were also poor in wave 4. Of those who were non-poor in wave 1, 79% were also non-poor in wave 4, while 21% transitioned into poverty between 2008 and 2014/2015. The probability of transitioning out of poverty over the four waves was therefore approximately five percentage points higher than the probability

[6]There is some sensitivity analysis to the choice of poverty line through the use of a measure of ‘severe’ poverty later in this chapter. Additional sensitivity tests using the StatsSA upper bound poverty line of R945 are available from the authors.

of transitioning into poverty over the same period for the balanced panel members.

Table 3.3: Transitions into and out of poverty across waves

		Wave 2				Wave 3	
		Poor	**Non-poor**			**Poor**	**Non-poor**
Wave 1	**Poor**	88.40	11.60	**Wave 2**	**Poor**	84.09	15.91
	Non-poor	26.48	73.52		**Non-poor**	20.26	79.74
		Wave 4				**Wave 4**	
		Poor	**Non-poor**			**Poor**	**Non-poor**
Wave 3	**Poor**	79.30	20.70	**Wave 1**	**Poor**	73.40	26.60
	Non-poor	20.74	79.26		**Non-poor**	21.36	78.64

Source: Own calculations from the first four waves of NIDS. Attrition-corrected panel weights applied to members of the balanced panel.

In Table 3.4 each cell within each panel gives the total proportion of balanced sample members in each transition state. The four cells in each panel sum to 100%, rather than each row summing to 100% as was the case in Table 3.3. Focusing on the shaded panel once again, we see that almost 54% of the sample of balanced panel respondents were poor in both wave 1 and wave 4. Just over 21% of respondents had real per-capita household incomes above R1 283 in both wave 1 and wave 4. Almost one fifth of respondents were poor at the start of the period, and non-poor at the end, while the opposite is true of 5.7% of the balanced panel.

Table 3.4: Poverty transitions: Proportion of sample by transition status

		Wave 2				Wave 3	
		Poor	**Non-poor**			**Poor**	**Non-poor**
Wave 1	**Poor**	64.69	8.49	**Wave 2**	**Poor**	60.37	11.42
	Non-poor	7.10	19.71		**Non-poor**	5.71	22.49
		Wave 4				**Wave 4**	
		Poor	**Non-poor**			**Poor**	**Non-poor**
Wave 3	**Poor**	52.41	13.68	**Wave 1**	**Poor**	53.72	19.47
	Non-poor	7.03	26.88		**Non-poor**	5.73	21.09

Source: Own calculations from the first four waves of NIDS. Attrition-corrected panel weights applied to members of the balanced panel.

Our final set of transition matrices, shown in Table 3.5, draws on the definition of 'severe' poverty used in Carter and May (2001), and classifies individuals as being in severe poverty if their real per-capita household income is less than half of the poverty line. Therefore the threshold for severe poverty in this context is R641.50, and the threshold for poverty is between R641.40 and R1 283 in January 2015 rands.

Of those who were in severe poverty in wave 1, 78.4% were either in severe poverty or in the 'poor' category in wave 4, implying that just over one fifth of the severely poor in wave 1 were non-poor in wave 4. The transition rates for those who were poor in wave 1 are higher when compared to the severely poor category, and this is to be expected as respondents could move in two directions if they were in the middle category at the beginning of the time period. Of those who were poor in wave 1, 27.5% transitioned down into the poorest category in wave 4, while just over 40% escaped poverty in the 2008 to 2014/2015 period. The non-poor/non-poor cell shows the highest level of stability, with 79% of respondents remaining non-poor in wave 4, conditional on being non-poor in wave 1. The proportions of non-poor wave 1 respondents transitioning into poverty or severe poverty by wave 4 are 13% and 8.5% respectively.

The final panel in the bottom left section of the table contains cells that sum to 100%. This allows us to see the overall proportion of respondents in each of the nine cells corresponding to different poverty transitions. 53.72% of the members of the balanced panel were in poverty or severe poverty in both wave 1 and wave 4.[7] This table highlights that most of those who were trapped in poverty were in fact trapped in severe poverty - 29% of all the balanced panel members were in this category. The proportion of the sample that was severely poor in wave 1 and non-poor in wave 4 stands at 11.5%, while 8% were poor in wave 1 and non-poor in wave 4. Just over one fifth of balanced panel members were non-poor in both waves, while about 6% transitioned from being non-poor into being either poor or severely poor.

[7]This corresponds to the proportion in the upper left cell of the shaded area in Table 3.4.

Table 3.5: Transitions with finer poverty levels

			Wave 2					**Wave 3**	
		Severe	**Poor**	**Non-poor**			**Severe**	**Poor**	**Non-poor**
	Severe	73.16	19.48	7.36		**Severe**	63.14	25.64	11.21
Wave 1	**Poor**	42.31	34.77	22.92	**Wave 2**	**Poor**	35.49	37.33	27.18
	Non-poor	12.25	14.23	73.52		**Non-poor**	8.30	11.95	79.74
			Wave 4					**Wave 4**	
		Severe	**Poor**	**Non-poor**			**Severe**	**Poor**	**Non-poor**
	Severe	60.21	24.36	15.43		**Severe**	53.88	24.55	21.57
Wave 3	**Poor**	33.35	36.87	29.78	**Wave 1**	**Poor**	27.47	32.48	40.04
	Non-poor	9.36	11.38	79.26		**Non-poor**	8.52	12.83	78.64

Note: In this panel the cells sum to 100%

			Wave 4	
		Severe	**Poor**	**Non-poor**
	Severe	28.69	13.07	11.49
Wave 1	**Poor**	5.48	6.48	7.98
	Non-poor	2.29	3.44	21.09

Source: Own calculations from the first four waves of NIDS. Attrition-corrected panel weights applied to members of the balanced panel.

Roughly 83% of balanced panel members are African, and although this group drives the overall results discussed in the previous three tables, it is interesting to highlight the results for African respondents in isolation. This analysis can be found in Tables 3.C.1, 3.C.2 and 3.C.3 in the appendix. A comparison of Table 3.3 to Table 3.C.1 shows that approximately three quarters of the full sample of balanced panel members (including Africans) were poor in wave 4 if they were poor in wave 1. The proportions transitioning out of poverty were therefore also approximately the same. There is a fairly large difference when comparing those who started off non-poor. 78.64% of the full sample who started off non-poor remained non-poor, but for Africans this proportion was lower at 71.52%. Although the transitions out of poverty were similar for African and non-African panel members, Africans were far more likely to transition into poverty than the rest of the sample. Of course, with Africans making up more than 83% of the balanced panel sample, the overall numbers are largely driven by this group. Comparing Africans with non-Africans (rather than with the full sample) reveals some starkly different numbers. Table 3.C.2 in the appendix shows that 60.36% of Africans were poor in wave 1 and wave 4. This compares to only 21.82% of non-Africans.[8] On the other hand,

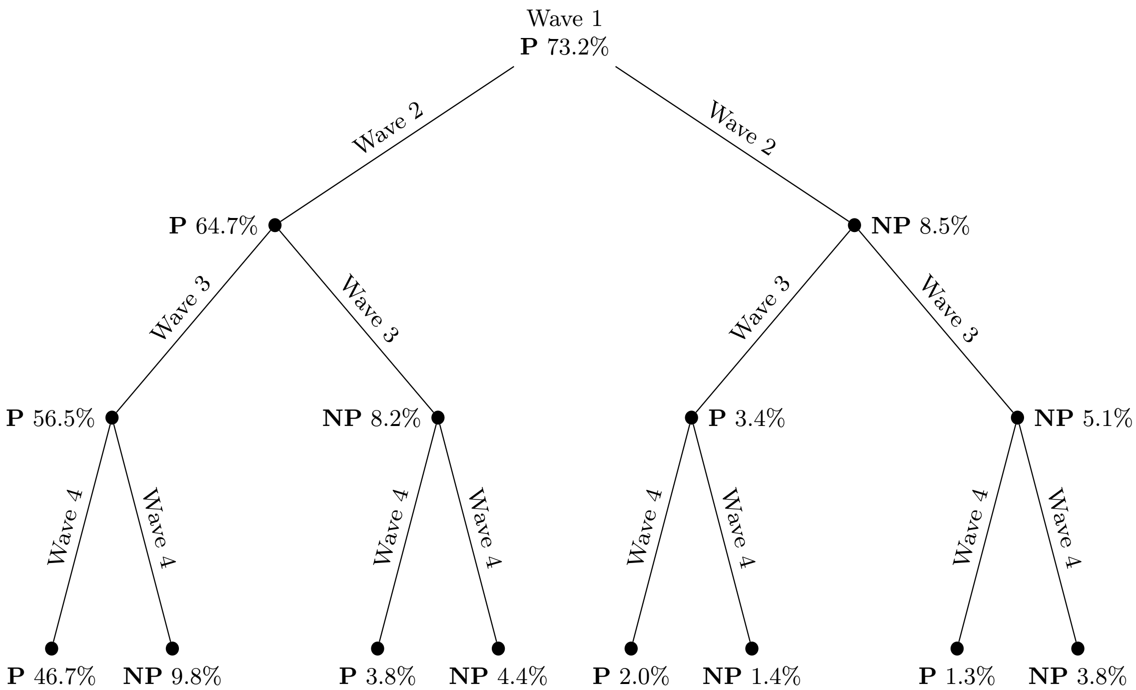
[8]This table is not shown in the appendix. Full tables of comparisons are available from the author.

56.58% of non-Africans were non-poor in both wave 1 and wave 4, while the corresponding number for African panel members was only 13.68%. Finally, as can be expected given the previous results, the proportion of Africans who were in severe poverty in both wave 1 and wave 4 was higher than for the rest of the balanced panel (see a comparison between Table 3.5 and Table 3.C.3). African respondents were more likely to transition from severe poverty into non-poverty (12.69% of the African sample experienced this transition), but were also more likely to transition from non-poverty into either poverty or severe poverty.

Poverty over four waves

Presenting all possible combinations of poverty status for balanced panel members across four waves is a significant challenge. There are 16 different possible states (PPPP, PPPN, PPNN,...,NNNN), compared to 8 different states if three waves are used, and 4 different states if two waves are used. The approach of presenting the different combinations of states of poverty and non-poverty over four waves as 16 possible paths can be found in Table 1 of Jarvis and Jenkins (1997), who provide all 16 possible combinations for the first four waves of the British Household Panel Survey. In this chaper a different approach is taken, and in Figure 3.1 and Figure 3.2 we use poverty transition trees to show the proportion of the balanced sample that was in each possible state over each of the four waves. The choice to show poverty transitions in this way is made because the trees contain more information than a table of 16 states, as the proportion of sample at each node is reported, rather than only the proportions who end up in each of the mutually exclusive final 16 categories. Showing the paths as a single table can be thought of as a special case of the transition tree approach, with the 16 terminal nodes in Figure 3.1 and Figure 3.2 representing the 16 possible final states. Each node of the tree represents an unfolding combination of possible states that a respondent could be in. For example, the top node in Figure 3.1 shows that 73.2% of balanced panel respondents fell below the poverty line of R1 283 per capita per month (P) in wave 1. Moving down a node, 64.7% of balanced panel members were poor in wave 1 and in wave 2 (PP). Moving down another node and going to the right this time, we see that 8.2% of balanced panel members were poor in wave 1, poor in wave 2, and non-poor in wave 3 (PPN). The terminal nodes show the final four wave combinations, along with the proportion of sample members in each. The PPPP node (the first terminal node in Figure 3.1) shows that 46.7% of balanced panel members were poor in each wave in which they were interviewed. Almost 10% of the sample was poor in the first three waves but exited poverty in the fourth wave. Going right from the initial node, we see that 8.5% of the balanced panel transitioned out of poverty between wave 1 and wave 2 (PN). Of that 8.5%, just over 5% remained non-poor in wave 3 (PNN), while 3.4% transitioned back into poverty between wave 2 and wave 3 (PNP). Finally, 3.8% of the balanced panel was in poverty in wave 1, but transitioned out of poverty in wave 2 and remained non-poor in all subsequent waves (PNNN).

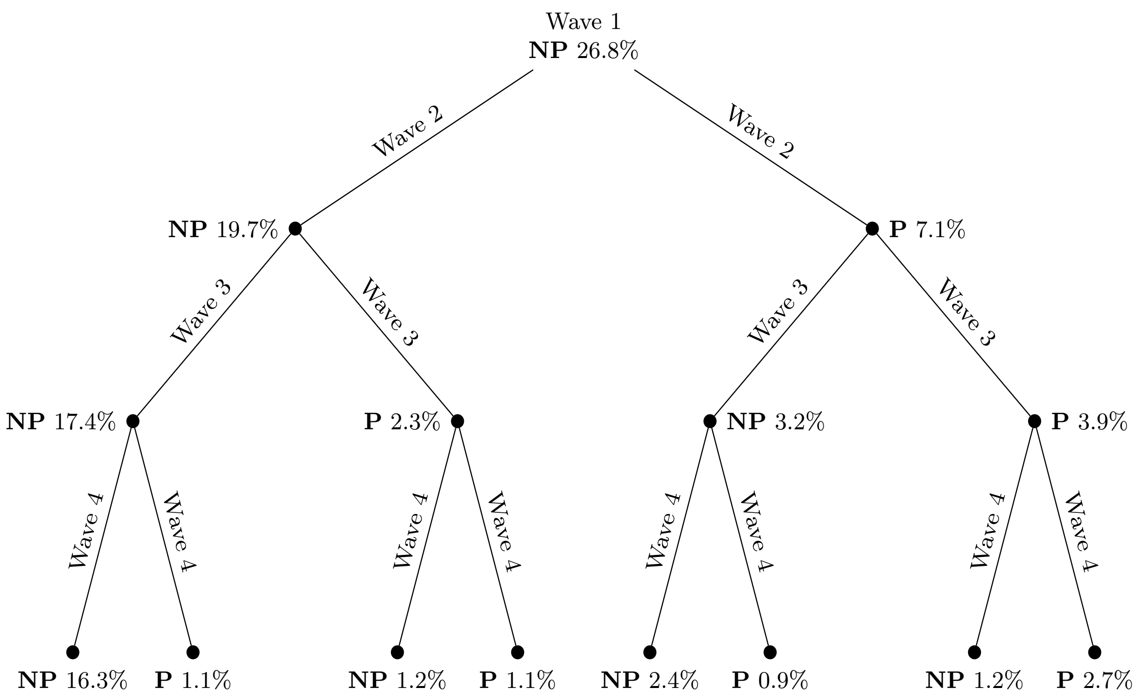
Figure 3.1: Poverty transition tree: Poor in wave 1



Source: Own calculations from the first four waves of NIDS. Attrition-corrected panel weights applied to members of the balanced panel. **P** = Poor, **NP** = Non-poor.

Figure 3.2 adopts the same approach, except this time for the 26.8% of balanced panel members who were non-poor in wave 1. The eight terminal nodes of this tree combined with the eight terminal nodes of the previous tree provide all 16 possible poverty transition states. The same goes for the eight possible states in wave three, and the four possible states in wave 2. In this figure we see that only 16.3% of all our balanced panel members were non-poor in each of the four waves. In fact, 7.1% of the total sample was non-poor in wave 1 and fell into poverty in wave 2 (NP). A slightly higher proportion of those on the NP path remained in poverty by wave 3 (3.9% at the NPP node) than exited poverty by wave 3 (3.2% at the NPN node). The terminal node of the far right of the tree shows that 2.7% of the balanced panel were non-poor in the first wave, but then feel into poverty and did not transition out in any subsequent wave (NPPP). Just under 1% of the sample transitioned at each wave, conditional on starting off non-poor (NPNP), and this is roughly the same proportion as those who started off poor and also transitioned in every wave (PNPN). 7.7% of the balanced panel sample started off non-poor in wave 1, but experienced one wave of poverty during either wave 2, 3 or 4 (calculated as the sum of NPNN, NNPN and NNNP). This is in contrast to 3.8% of balanced panel members who were in poverty only in wave 1 (PNNN).

Figure 3.2: Poverty transition tree: Non-poor in wave 1



Source: Own calculations from the first four waves of NIDS. Attrition-corrected panel weights applied to members of the balanced panel. **P** = Poor, **NP** = Non-poor.

Although the unconditional poverty rates of the balanced panel in each wave are of less interest to us than poverty transitions between waves, they are nevertheless embedded in the transition trees, and can be calculated by summing the percentages at each ‘P’ node. Doing so reveals the unconditional poverty rates as 73.2%, 71.8%, 66.1% and 59.5% in each of the first four waves respectively.[9]

Another way of displaying poverty states across four waves is presented in Table 3.6 which shows the number of times respondents were recorded as being in poverty and severe poverty over the total time period. Although this is a simpler display of poverty states, we sacrifice the ability to show every possible state over every wave, as was done in the previous two figures. As we have already seen, only 16.28% of the sample of balanced panel respondents were classified as non-poor in every wave in which they were observed. This is in stark contrast to the 46.7% of respondents who were living below the poverty line of R1 283 per month in each of the four waves in which they were interviewed. The proportion of respondents who were in severe poverty (living on less than half the poverty line) ranges from 29% who were never recorded as being in severe poverty to 16% who were recorded as being in severe poverty in a single wave. Just over 18% of respondents were recorded as being in severe poverty in all four waves, though 55.3% experienced severe poverty in at least half the waves in which they were interviewed. The middle column of the table presents the proportion of African balanced panel members who were poor between zero and four times over the first four waves of NIDS. As expected, given the analysis of the transition matrices earlier in the chapter, dynamics in the African subsample drive the overall findings. Although just over 16% of the full sample did not experience poverty in any wave, the corresponding proportion for African respondents was just under 9.5%. The proportion of African respondents experiencing poverty in either 2 or 3 waves lines up relatively closely with the overall numbers, but significantly more Africans were poor in every wave compared to the full sample (53% versus 46.7%). Perhaps a more telling comparison is between African and non-African respondents. As already noted, only 9.5% of African sample members were non-poor in all four waves of NIDS. The proportion in this category amongst non-African sample members was close to 50%. At the other extreme, over one fifth of African respondents were in severe poverty every time they were observed, while the corresponding proportion for non-Africans was a little under 4%.

[9] The unconditional poverty rates of balanced panel members should not be treated as being representative of national poverty rates, which can be calculated by using each separate wave as a cross-section.

Table 3.6: Number of times observed in poverty between 2008 and 2014/2015

	Overall		**African**		**Non-African**	
	Poverty	**Severe poverty**	**Poverty**	**Severe poverty**	**Poverty**	**Severe poverty**
0	16.28	28.81	9.48	20.77	48.89	67.38
1	8.51	15.89	7.79	16.2	11.98	14.37
2	10.30	17.74	9.86	19.23	12.42	10.61
3	18.24	19.41	19.8	22.67	10.73	3.76
4	46.67	18.15	53.07	21.13	15.98	3.88

Source: Own calculations from the first four waves of NIDS. Attrition-corrected panel weights applied to members of the balanced panel.

Chronic versus transitory poverty

We have seen that a large proportion of the sample remained in poverty in all four waves. It would be useful to complement this finding by considering the extent to which the overall poverty rate is made up of chronic versus transient poverty in each of the four waves.

This is, in some ways, an extension of the two-period conception of chronic poverty in South Africa that is presented in Carter and May (2001). The authors find that 18% of African households were in chronic poverty in KwaZulu-Natal between 1993 and 1998, and reinforce the fact that a large proportion of South Africans were unable to take advantage of the early post-apartheid economy.

Carter and May (2001), focus in particular on the role of assets in driving poverty over time using the KIDS dataset. The authors use a combination of money metric and asset measures in order to classify households as being either stochastically poor or structurally poor. von Fintel et al. (2016) use the framework developed by Carter and May (2001) in order to investigate child poverty in general, and chronic child poverty in particular in South Africa. Other studies that deal *inter alia* with chronic poverty in South Africa (also using the KIDS dataset) are Roberts (2001) who delineates chronic and transitory poverty by demographic characteristics, and Aliber (2003) who discusses chronic poverty in light of some of the macroeconomic strategies adopted by the South African government in the late 1990s.

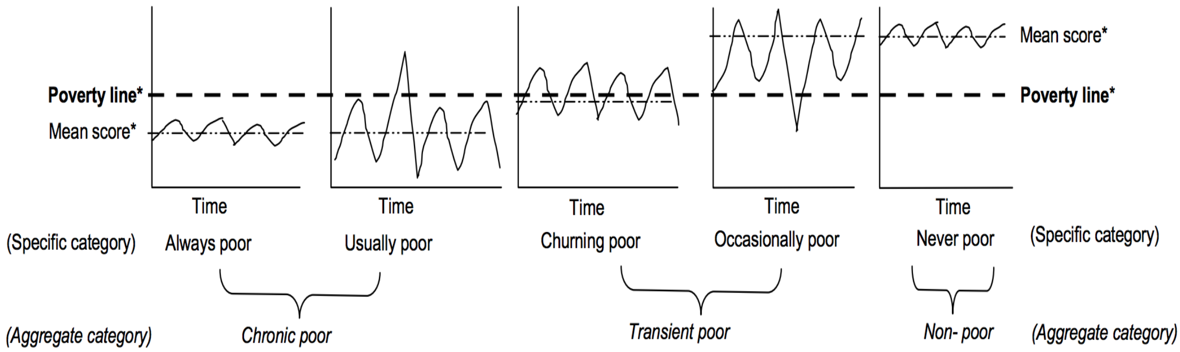
Given that we have four waves of data to work with, we can characterize the poverty-time interaction in a number of ways. We follow Hulme and Shepherd (2003) who conceptualise different five types of poverty in an adaptation of the methodology found in Jalan and Ravallion (2000). The original Jalan and Ravallion (2000) study of chronic versus transient poverty in China uses longitudinal data over a six year period. The transient component of poverty in that study is thought of as the contribution of inter-temporal variability in living standards to poverty, while the chronic (or non-transient) component is thought of simply as time mean consumption/income for all dates. The authors use the squared poverty gap as their poverty measure and find that just under half of poverty in China can be explained by the transient component. However, only 6% of individuals lived in households which were persistently poor, and 54% were classified as never-poor. The five different characterisations of poverty over time are applied to NIDS in the following way.

- Always poor: household income per capita measures are below the poverty line in all four waves of NIDS.
- Usually poor: mean household income per capita over the four waves of NIDS is less than the poverty line, but the panel member is not poor in every period.[10]
- Churning poor: mean household income per capita over the four waves of NIDS is in the neighbourhood of the poverty line[11] but the panel member is sometimes poor and sometimes non-poor in different periods.[12]
- Occasionally poor: mean household income per capita over the four waves of NIDS is above the poverty line, but the respondent is poor in at least one wave.
- Never poor: household income per capita is above the poverty line in all waves of NIDS.

These five characterisations of poverty are represented graphically in Figure 3.3:

[10] This is effectively the same definition used by Jenkins (2011) in defining chronic poverty.
[11] In this study we choose a window of 10% below the poverty line.
[12] This can be thought of as a special case of 'usually poor' with the additional restriction being that the respondent's average household income per capita is close to the poverty line.

Figure 3.3: Five different characterisations of poverty over time

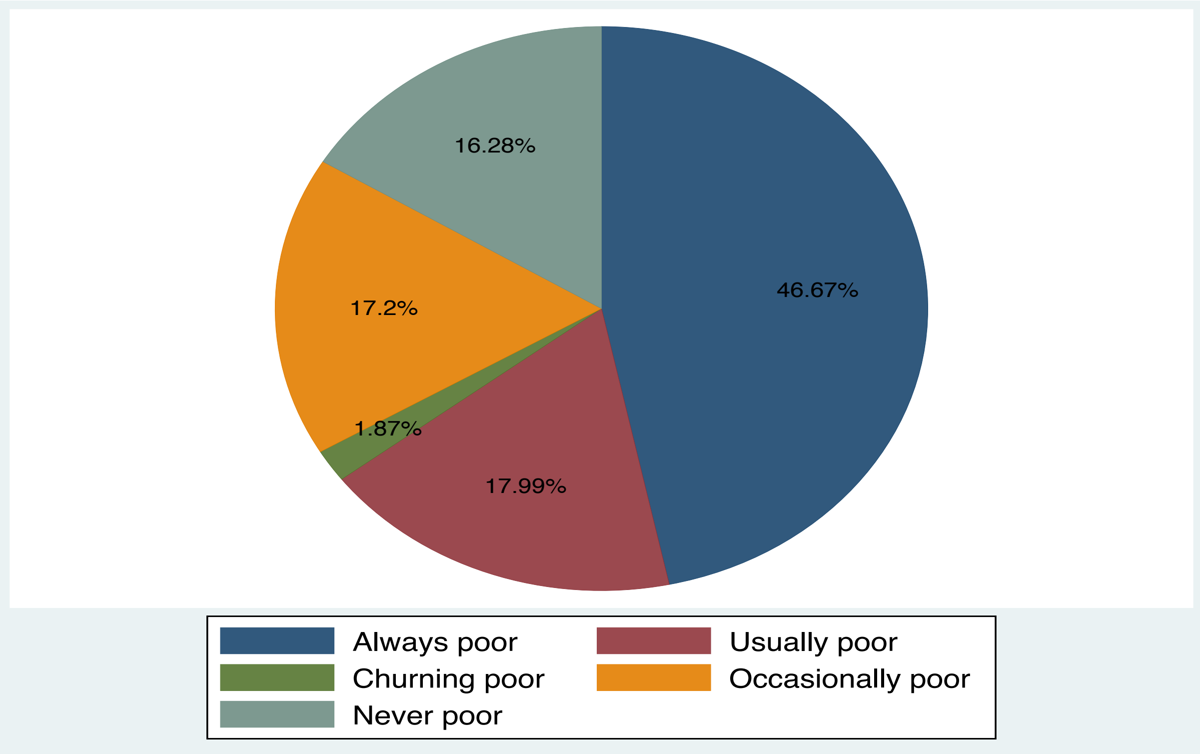


Source: Reproduced from Hulme and Shepherd (2003).

In Figure 3.4 we present the five different poverty states for members of the NIDS balanced panel. The single largest group is the 46.67% of panel members who were poor in all four waves. This proportion, added to the 18% of respondents who were usually poor, gives a 'chronic poverty' percentage of almost two-thirds over the period of 2008 to 2014/2015. By far the smallest group is those who we have defined as 'churning' poor - their average household income per capita was within 10% below the poverty line, but who were non-poor in at least two waves. The second category comprising 'transient poverty' is made up of those panel members whose average household income per capita was above the poverty line, but who were poor in at least one wave. These made up 17.2% of respondents, meaning that almost one fifth of the balanced panel members were in the 'transient poor' category. Finally, only 16.28% of respondents were non-poor in every single wave.[13]

[13]This corresponds to the number presented in the leftmost terminal node of Figure 3.2.

Figure 3.4: Types of poverty experienced by the balanced panel



Source: Own calculations from the first four waves of NIDS. Attrition-corrected panel weights applied to members of the balanced panel.

3.5 Trigger events associated with movements into and out of poverty

Demographic versus income events

Given that our poverty line is a threshold of real monthly household income per capita, we can expect changes either through the numerator (income events) or through the denominator (demographic events). The trigger events that we use in this chapter are a combination of those found in Jenkins (2011) and Woolard and Klasen (2005), both of which are based on the original exposition in Bane and Ellwood (1986). The first kind of demographic event is a change in the household head and/or a change in the composition of the household. This is typically one or more people entering/leaving the household due to birth, migration or death. Thus the dynamics of household composition affects our sample, even though the people entering/leaving the household may not be balanced panel members.[14] The first category of 'head or composition changed' therefore includes headship changes as well as other household formation changes.

The second category is assigned to cases where the head of the household did not change in between waves, but the composition of the household did. Given that the head did not change, it was necessary to determine whether changes in needs outweighed changes in income. This was accomplished by comparing the proportional change in the household size for each individual compared to a proportional change in total household income, following Jenkins (2011). If the proportional change in needs was larger (in absolute terms) than the proportional change in income, then the trigger event 'needs > money' was assigned to all individuals in the relevant household.

There is, of course, a link between demographic events and income events affecting the welfare of household members. This takes place primarily through the migration of household members. Changes in the composition of the household may themselves be driven by changes in income sources, as some research on the role of the South African old age pension on labour supply suggests (Posel and Casale (2003), Posel et al. (2006) and Ardington et al. (2009) find positive labour supply effects, while Bertrand et al. (2003), Ranchhod (2009) and Sienaert (2008) find negative effects). Although this form of physical mobility is not the focus of the chapter, it is worth highlighting some key findings about migration that have been uncovered using NIDS data. The literature on physical migration and mobility in South Africa is very thin, as noted in Posel (2010), which uses NIDS data to investigate some of the correlates of physical mobility in the country. The paper compares findings from 1993 PSLSD data to 2008 NIDS data and concludes that a larger proportion of those who migrated around 2008 did so for employment reasons, compared to those who migrated around 1993. Another, perhaps surprising, finding is that although migration was driven by perceptions about labour market access, the connection to origin households via remittances was weaker in the NIDS data than in the PSLSD data. Clarke and Eyal (2014) use the first two waves of NIDS to study migration and find that receipt of the state old age pension and, to a lesser extent, receipt of the child support grant is negatively associated with the probability of migration of co-resident, non-eligible

[14]Each wave of NIDS collects data on all members of the household in which panel members reside, whether or not these members were part of the base wave in 2008.

adults in the household. They also find an inverse-U relationship between the probability of migrating and household income. Respondents aged between 18 and 30 are the most likely to migrate, while those with school-aged children and those living in rural areas are less likely to move.

There are five types of income trigger events. The first three are: changes in formal earnings of the household head, formal earnings of the spouse of the household head, and formal earnings of any other household members. The final two income triggers are changes in remittance income received by the household, and changes in income from government grants received by the household.[15] Income events are ranked according to the size of the change between waves. So, for example, if the household head's real formal earnings fell by R200, the spouse's real formal earnings fell by R800 and there was no change in the other income triggers, then the appropriate trigger event is 'fall in spouse's formal labour market earnings'. Finally, there is an 'inconclusive' category which indicates households in which no clear ranking can be established.

More formally (and assuming that all trigger events are assigned), we borrow notation from Jenkins (2011) to show that the probability of exiting poverty[16] is made up of mutually exclusive events 1 to J.[17]

$$Pr(\text{exit poverty}) = \sum_{j=1}^{J} Pr(\text{exit poverty via trigger j}) \tag{3.1}$$

Given that each event 1 to J can be formulated as the product of the probability of poverty exit, conditional on event j, and the probability of event j itself occurring, we have:

$$Pr(\text{exit poverty}) = \sum_{j=1}^{J} Pr(\text{exit poverty}|\text{trigger j}) \times Pr(\text{trigger j}) \tag{3.2}$$

[15] The state old age pension and the child support grant make up by far the largest share of household income from state grants. The pension is means tested and is paid to eligible recipients who are aged 60 and above. It is a relatively high amount at about 1.75 times the median of income for African respondents (Clarke and Eyal, 2014). The child support grant is also means tested and is paid to the primary caregiver of the child until the child reaches 18 years old. The amount of the child support grant is approximately one quarter of the amount of the old age pension.

[16] The notation for the probability of entering poverty via trigger k is easily seen from this example.

[17] Jenkins (2011) and Jenkins and Rigg (2001) also provide estimates that do not assume mutually exclusive trigger events, though these do not form part of this chapter.

It is important to note that although this analysis of demographic versus income events is interesting and useful, we should be very cautious about assigning causality from the trigger to the transition. As Jenkins (2011) notes, it is tempting to say that losing an employed member caused a particular household to enter poverty, but it could also be the case that a household entered poverty first, and this stress caused the household to break up.

The role of trigger events

The first feature to note about Table 3.7 is the fact that demographic events were more heavily weighted than income events in terms of their importance in explaining transitions both into and out of poverty during the period under study. A demographic change in the household was the main trigger for 56% of individuals who entered poverty between wave 1 and wave 4. A fall in the real formal labour market earnings of the household head was the primary income correlate of entering poverty. This was the primary trigger for poverty entry for about one fifth of those who entered poverty between wave 1 and wave 4. Falling earnings for household members who were not the household head or spouse of the household head triggered poverty entry for between 10% and 14.5% of balanced panel members, depending on which transition is the focus. The shares of falling remittances and falling income from government grants were relatively similar for poverty entry during each of the three time periods under study, and were generally the relevant triggers for between 3% and 5% of balanced panel members.

For those respondents who exit poverty, the head change/compositional change share from wave 1 to wave 4 is almost 14 percentage points higher than its counterpart in the poverty entry category. It is interesting that the needs $>$ money category (no change in the household head but a compositional change in the household) contributes relatively little to the total explanation of poverty exit – dropping to as low as 0.4% for the wave 1 to wave 3 transition.

The income triggers tell quite a different story for poverty exit than they do for poverty entry. An increase in the earnings of the household head is the main poverty exit trigger for almost one quarter of those who left poverty between wave 1 and wave 2, but its importance falls to only 4% for the full wave 1 to wave 4 period. The importance of the earnings of the spouse of the household head is relatively muted over the whole period, reaching a high of 3.9% for the wave 2 to wave 3 transition. An increase in labour market earnings from household members who are not the head or married to the head is the main poverty exit trigger for about 10% of balanced panel members for the wave 1 to wave 4 transition. This share is similar to its counterpart in the poverty entry panel. The importance of increased remittance income is fairly muted for poverty exit triggers, as it was for poverty entry triggers. One significant difference between wave 1 to wave 4 poverty entry and poverty exit triggers is the role of income from government grants. A drop in grant income was the main poverty entry trigger for only 3.5% of those who entered poverty. In stark contrast, an increase in income from government grants was the main trigger precipitating poverty exit for 23% of balanced panel members. This is a reflection of both the success of the targeting and expansion of the state's grant system, and the failure of the labour market to act as the main driver of poverty reduction in the country.[18]

Table 3.7: Trigger events associated with poverty entry and exit

	Poverty entry				Poverty exit			
	W1 to W2	**W2 to W3**	**W3 to W4**	**W1 to W4**	**W1 to W2**	**W2 to W3**	**W3 to W4**	**W1 to W4**
Demographic								
Head changed	34.83	49.49	52.02	42.10	34.34	47.50	50.55	55.91
Needs > money	11.70	6.75	12.96	13.73	3.75	0.62	0.37	2.64
Demographic share	**46.53**	**56.24**	**64.98**	**55.83**	**38.09**	**48.12**	**50.92**	**58.55**
Income								
Head labour earnings	18.86	15.72	10.02	19.11	23.57	16.99	4.60	4.00
Spouse labour earnings	4.64	1.75	2.82	3.15	2.70	3.86	3.59	1.38
Other labour earnings	14.47	12.45	9.56	10.60	17.58	13.65	10.39	10.19
Remittances	4.67	3.65	3.91	3.98	2.18	5.08	4.50	2.10
Grant income	4.52	3.31	2.26	3.53	9.89	7.39	23.97	23.16
Income share	**47.16**	**36.88**	**28.57**	**40.37**	**55.92**	**46.97**	**47.05**	**40.83**
Inconclusive	6.32	6.88	6.44	3.80	5.99	4.91	2.02	0.62
Total	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
Observations	963	925	1 266	804	1 317	1 937	2 324	3 228

Source: Own calculations from the first four waves of NIDS. Attrition-corrected panel weights applied to members of the balanced panel.

Woolard and Klasen (2005) also analyse demographic versus income events in triggering poverty exit. Their focus is on African households, and they use two waves of data from 1993 and 1998. The sample sizes are relatively small, with 129 households entering poverty and

[18]Eyal and Burns (2016) show the rapid growth in the reach and effect of the child support grant over the first three waves of NIDS. By the third wave of NIDS 89% of those who were age and income eligible to receive the child support grant actually received it. This equates to about 59% of all household in the third wave of NIDS. A report by the department of social development shows that the number of children covered by the child support grant increased from 2 million in 2002, to 8 million in 2008, to 11 million in 2011 (DSD et al., 2012). Evidence of the effective targeting of the countrys state old age pension can be found *inter alia* in Abel (2013) and Standish-White and Finn (2015).

223 exiting poverty over the two waves. In contrast to the findings in this chapter, Woolard and Klasen (2005) attribute most of the transitions into and out of poverty to income, rather than demographic events. They find that demographic events are responsible for 27.4% of households falling into poverty, and 23.6% of households exiting poverty. This is far less than this chapter's corresponding figures of 55.8% and 58.6% for poverty entry and exit respectively. Of income events between 1993 and 1998, Woolard and Klasen (2005) find that the single most important factors for households entering poverty are the head of the household losing a job, followed by another household member losing a job, followed by a drop in remittances. Income events most strongly associated with poverty exit are another household member gaining employment, followed by the household head gaining employment, followed by a rise in remittances.

The prominence of demographic trigger events over income trigger events in explaining paths into and out of poverty may be driven in part by the choice of the per capita equivalence scale. In order to investigate how sensitive the results are to the choice of the equivalence scale, Table 3.D.1 in the appendix presents the demographic and income trigger events for the same equivalence scale used in Woolard and Klasen (2005). That is, total household income is divided by $(adults + 0.5 \times children)^{0.9}$, rather than simply by household size. Note that in comparing Table 3.7 and Table 3.D.1 we are not comparing exactly the same subsamples, as the choice of equivalence scale has an effect on which households transition into and out of poverty. The first thing to note about the use of the new equivalence scale is that the number of observations falling into poverty between wave 1 and wave 4 is lower than if household income is divided by the number of household members (479 compared to 804). The effect works in the same way when considering those who exited poverty (5 945 using the new equivalence scale compared to 3 228 using the old equivalence scale).

The overall effect of using the modified equivalence scale changes the relative weightings of the demographic and income effects differently whether one considers poverty entry or poverty exit. For the former, the share of demographic trigger events in explaining poverty entry rises from 55.83% to 58.92% when analysing transitions between wave 1 and wave 4. However, this is the only transition in which the demographic share rises – it falls for the wave 1 to wave 2, wave 2 to wave 3 and wave 3 to wave 4 transitions. This increase is entirely due to the rise in the importance of a change in the head of the household, rather than a change in the needs to money ratio. Continuing with the wave 1 to wave 4 poverty entry transitions, the importance of a negative change in the labour market earnings of a household member who was not the head, or married to the head, decreases substantially if the modified equivalence scale is used. This change has the opposite effect on the importance of a change in grant income in the household, with the share of trigger events attributable to this factor rising from 3.53% to 7.38%.

The dynamics underlying poverty exit are also affected by a change in the underlying equivalence ratio, but the qualitative result of demographic events outweighing income events remains the same. If a per capita equivalence scale is used, then 58.55% of poverty exits can be explained by demographic events. The corresponding share if the modified equivalence scale is used is slightly lower at 54.26%. The fall in the demographic share is due to lower weighting of both components of that measure. The change in equivalence scale raises the shares of all the items in the income change category except for changes in grant income, which falls very slightly from 23.16% to 22.71%. In summary, the choice of whether to use a per capita equivalence scale or the equivalence scale used in Woolard and Klasen (2005) makes very little difference to the qualitative findings of the trigger events approach to explaining transitions into and out of poverty.

3.6 A Markovian model of poverty transitions

We now change the focus of the chapter from descriptive statistics to a careful modelling of the dynamics. Although it may be tempting to adopt a univariate probit approach to modelling poverty transitions, this approach may produce biased results by not controlling for initial conditions and selective attrition. In this section we model transitions while specifically controlling for both of these factors. The model exposition and implementation closely follows Cappellari and Jenkins (2004) and Jenkins (2011).

As noted by Jenkins (2011), much of the international literature on poverty dynamics adopts either a hazard model approach or a variance-component approach to understanding transitions. A third approach, and one that has not been applied very often in the international literature nor, indeed, at all to South African data, is a so-called first order Markov model of poverty dynamics.

The Markovian approach to poverty dynamics has, as a first advantage over alternative methods, the ability to take initial conditions (the presence of a lagged dependent variable in the model that is correlated with unobserved heterogeneity) and non-random attrition into account. The approach, in theory, allows the researcher to overcome the biases presented by both the initial conditions and non-random attrition by modelling both jointly with the probability of transitioning into or out of poverty. This is not something that can easily be incorporated into a more traditional hazard model of poverty dynamics.

The initial conditions problem in the analysis of transitions was first suggested by Heckman (1981*a*). Essentially, the problem is that if we find a certain level of state dependence when analysing poverty, it may be that those who are more likely to be permanently poor are over-represented in the sample. Another way of thinking about the problem is that the start of the period of analysis (in our case wave 1) does not coincide with the start of the process that generates poverty or non-poverty outcomes. As noted in Arulampalam et al. (2000),[19] a model may control for unobserved heterogeneity, but it is also important to separate out the effect of state dependence from unobserved heterogeneity. That is, the initial condition must be modelled explicitly as it may be correlated with the unobservables.

A second advantage of the approach is that it allows for left-censored poverty spells to be incorporated into the model. A conventional hazard model would drop all data for respondents who are poor in every period (almost half our sample), and data for those who are non-poor in every period (a further 16% of our sample). This means that a lot of individuals would fall out of the dataset, which increases the probability of the estimation sample being unrepresentative (Jenkins, 2011). The cost of this second advantage relative to other methods is that it is gained by the assumption that there is no duration dependence.

A third advantage that this model offers over its alternatives is the ability to circumvent the strict exogeneity assumption made about the explanatory variables. In other words, as noted in Biewen (2009), a Markovian approach to poverty transitions avoids having to make

[19]In the context of unemployment dynamics.

the assumption that there is no feedback from dependent variables on future values of the explanatory variables. The cost of doing so is that this approach will yield inefficient estimates due to the fact that the longitudinal nature of the data is not fully exploited.

In summation, the use of a Markovian model to analyse poverty dynamics in South Africa allows us to buy identification while accounting for unobserved heterogeneity, non-random attrition and initial conditions in a single framework, by making some strong distributional and exogeneity assumptions. The estimates obtained can be used to predict poverty spell lengths for individuals with different characteristics.[20]

The data suggest that controlling for initial conditions and selective attrition when modelling poverty dynamics in South Africa is warranted.[21] This is motivated by the output in Table 3.8 which shows poverty transitions for the pooled sample of respondents over the four waves of NIDS. It is important to note that this table is different to Tables 3.3, 3.4 and 3.5, as our modelling strategy requires us to used pooled data to model transitions, and so we do not restrict ourselves to the balanced panel in any future analysis in the chapter.

The first panel of the table presents poverty transitions for all respondents for whom household income was recorded in two consecutive waves. Just over 29% of those who were non-poor in year $t-1$ were poor in year t. Of those who were poor in year $t-1$, 14% were non-poor in year t. This panel confirms the findings in the descriptive section of this chapter by showing how the probability of being poor in a given year is highly dependent on the probability of being poor in the previous time period. Clearly, initial conditions are important here, and state dependence should not be ignored. Indeed, the poverty rate for those who were poor in year $t - 1$ is almost 57 percentage points higher than it is for those who were non-poor in the same year.

This 'naïve' transition matrix presents poverty transitions without controlling for individual and household heterogeneity, and may be thought of as reflecting 'aggregate' state dependence, something to which we will return. The estimation strategy that follows allows us to control for the determinants of initial poverty status, and allows for these determinants to be correlated

[20]One of the downsides of modelling poverty under the assumption of a first-order Markov process is that the model is not fully efficient as it only uses data from $t - 1$ and t.

[21]A full summary of the descriptive statistics for the observations used in the Markovian model can be found in Table 3.E.1 in the appendix.

to current poverty status. This allows us to uncover a measure of 'genuine' state dependence - the measurement of which we will return to after the results of the model have been presented.

While panel a) highlights the importance of initial conditions in determining poverty transitions, panel b) shows that ignoring selective attrition could be problematic at the estimation stage. The final column in Table 3.8 shows the extent of attrition between $t-1$ and t for non-poor and poor sample members respectively. The rate of attrition from $t-1$ to t differs substantially between the initially non-poor and the initially poor, and is 23.4% and 16.5% respectively. The relatively higher rate of attrition amongst non-poor sample members compared to poor sample members may result in a selected sample that biases our estimation of poverty dynamics.[22]

An overview of the nature of attrition across the first four waves of NIDS can be found in Chinhema et al. (2016), while Baigrie and Eyal (2013) contains a more detailed analysis of the determinants of attrition between the first two waves of NIDS. Chinhema et al. (2016) notes that non-contact[23] was the primary reason for respondents dropping out between waves 1 and 2, and waves 3 and 4. The biggest reason for attrition between wave 2 and 3 was the refusal of respondents to participate. The share of attrition attributable to respondents dying between waves was stable at around 15% for all three wave-to-wave transitions. Attrition amongst African respondents was mainly driven by non-contact, while the dominant reason for coloured and white respondents dropping out of the sample was refusal to participate. The overall attrition rate dropped from 22% to 16% to 14% for each successive wave-to-wave transition. This, however, hides stark differences in the attrition rates of each racial group. More than 50% of white respondents attritted in each new wave, while the rate of attrition of coloured respondents dropped from 27% to 18% to 16%. The African attrition rate decreased significantly with each additional wave of data - from 19% to 13% to 11%.

[22]Modelling non-random attrition in this way is different to correcting for attrition using panel weights, (as was done in the descriptive sections of this chapter) because the model explicitly takes account of the unobservables.

[23]By this we mean the inability of enumerators to locate the panel member.

Table 3.8: Poverty transitions with and without missing data

Poverty status at t-1	**Poverty status at t**		
	Non-poor	**Poor**	**Missing**
a) Sample with non-missing income at t			
Non-poor	70.86	29.14	
Poor	14.08	85.92	
Total	26.48	73.52	
b) All respondents			
Non-poor	54.28	22.32	23.40
Poor	11.75	71.74	16.51
Total	21.68	60.20	18.12

Source: Own calculations from the first four waves of NIDS. Pooled transitions with sample size of 40 850 individuals (panel b), and 88 090 person-waves. Rows sum to 100%.

Our estimation strategy in the chapter allows us to model poverty transitions while at the same time accounting for initial conditions and non-random attrition. The following section provides an outline of the theory underlying the empirics for our estimation of poverty dynamics.

Theoretical background to the estimation

The Markovian approach to poverty transitions models dynamics from base period $t-1$ to the next period t. There are four components to the model which are:

- The determination of poverty status in $t-1$ to account for initial conditions.
- The determination of whether monthly household per capita income is observed in periods $t-1$ and t, in order to account for selective attrition.
- The determination of poverty status in period t.
- The correlation between unobservables that influence each of the three processes above.

Initial poverty

In the base year there is a latent propensity for poverty over individuals $i = 1, ..., N$, individual and household explanatory variables in the vector x_{it-1}, parameters β and error term $u_{it-1} = \mu_i + \delta_{it-1}$ which is distributed N(0,1) and contains an individual-specific component plus an orthogonal white noise component:

$$p^*_{it-1} = \beta' x_{it-1} + \mu_i + \delta_{it-1} \tag{3.3}$$

As we only observe discrete outcomes of this latent model, we define a poverty indicator variable $P_{it-1} = 1$ if $p^*_{it-1} > 0$ (where 0 is the unobserved threshold) and zero otherwise.

Retention

We now move to model the probability that a respondent is observed in both the base and the subsequent wave of data. The latent propensity for retention, r^*_{it}, is given by the relationship:

$$r^*_{it} = \psi' w_{it-1} + \eta_i + \xi_{it} \tag{3.4}$$

where the error term v_{it} is once again distributed N(0,1) and is composed of an individual fixed effect η_i and a normal white noise error term ξ_{it}. R_{it} is a binary indicator for whether the respondent is observed in both periods, and is assigned a value of 1 if $r^*_{it} > 0$ where, once again, the threshold has been normalized to 0.

Current poverty

The third part of the model is the determination of poverty in period t. Once again, we adopt a latent variable approach with poverty status in period t being characterized by the following relationship:

$$p^*_{it} = \left[(P_{it-1})\, \gamma'_1 + (1 - P_{it-1})\, \gamma'_2 \right] z_{it-1} + \tau_i + \zeta_{it} \tag{3.5}$$

where γ_1, γ_2 and z_{it-1} are vectors and the composite error term ϵ_{it} is distributed N(0,1) and once again comprises individual (τ_i) and white noise (ζ_{it}) components. The vector of covariates z_{it-1} contains individual and household characteristics, as well as a constant term. Finally, let the observed poverty status $P_{it} = 1$ if $p^* > 0$ and zero otherwise. This is, of course, only observed if $R_{it} = 1$.

The specification above allows not only for base characteristics to impact poverty in the final period, but also for a characteristic to have a differing impact on the probability of entering or exiting poverty.

The three error terms, u_{it-1}, v_{it} and ϵ_{it}, are assumed to be distributed trivariate standard normal. There are three important correlations that we will estimate in order to parameterize the unobserved heterogeneity in the model. These are:

$$\rho_1 \equiv corr(u_{it-1}, v_{it}) = cov(\mu_i, \eta_i) \tag{3.6}$$

which gives the relationship between unobserved heterogeneity determining poverty in the base year and the probability of remaining in the sample. In this case, a positive correlation implies that poor respondents at $t-1$ are less likely to have attritted by period t.

$$\rho_2 \equiv corr(u_{it-1}, \epsilon_{it}) = cov(\mu_i, \tau_i) \tag{3.7}$$

which gives the relationship between unobserved heterogeneity that impacts on base year and final year poverty status. If this correlation is positive then it implies that respondents who started off poor are more likely to be poor in the next period than those who started off non-poor.

$$\rho_3 \equiv corr(v_{it-1}, \epsilon_{it}) = cov(\eta_i, \tau_i) \tag{3.8}$$

which is the relationship between the unobservables determining the probability of being retained in the sample and poverty status in the final period. In this case a positive correlation implies that those who are retained in the sample are more likely to remain poor or transition into poverty than those who attrit between the two time periods.

In any event, there are some interesting testable relationships to consider. If $\rho_1 = \rho_2 = \rho_3 =$

0, then poverty dynamics can be estimated separately using any univariate binary dependent variable model, such as a probit. However, if $\rho_1 = \rho_3 = 0$ then the process of attrition is ignorable and the model to be estimated becomes a bivariate probit. Finally, if $\rho_1 = \rho_2 = 0$ then we do not have to take initial conditions into account, and the poverty status in the base period may be treated as exogenous.

Given the descriptive statistics that have been presented, it is likely that the probability of the unobserved factors being uncorrelated with one another is very small. Incorporating initial conditions, non-random attrition and unobserved individual heterogeneity into the model requires the estimation of a partial likelihood estimator of the type used in Jenkins (2011) and Cappellari and Jenkins (2004).

From panel b) in Table 3.8 we see that there are six possible outcome combinations for a given sample member between $t-1$ and t. These are, non-poor to non-poor, non-poor to poor, poor to non-poor, poor to poor, non-poor to missing, and poor to missing. The setup of the model implies the following equations for poverty persistence (poor in both time periods) and poverty entry (non-poor in the first period, poor in the second) respectively:

$$s_{it} \equiv Pr(P_{it} = 1 | P_{it-1} = 1) = \frac{\Phi_2(\gamma_1' z_{it-1}, \beta' x_{it-1}; \rho_2)}{\Phi(\beta' x_{it-1})} \tag{3.9}$$

and

$$e_{it} \equiv Pr(P_{it} = 1 | P_{it-1} = 0) = \frac{\Phi_2(\gamma_2' z_{it-1}, -\beta' x_{it-1}; -\rho_2)}{\Phi(-\beta' x_{it-1})} \tag{3.10}$$

In the two equations above Φ_2 and Φ refer to the CDFs of the bivariate and univariate standard normal distributions respectively. Note that the regressors in these transition probabilities are measured using data from period $t-1$. This allows us to calculate transitions (or persistence) for those respondents with $R_{it} = 0$.

The sample log likelihood function contains six possible outcomes based on poverty status in the initial wave and on sample retention. Sample members who were retained in the panel fall into four possible outcome categories, depending on initial poverty status, while those who attritted are only observed as either non-poor or poor in the initial period. Therefore the log likelihood contribution of individual i whose poverty status is observed in the initial period is given by the following sample log likelihood function:

$$
\begin{aligned}
logL_i = {} & P_{it-1}R_{it}log\left[\Phi_3(k_i\gamma_1' z_{it-1}, m_i\psi' w_{it-1}, q_i\beta' x_{it-1}; k_i m_i\rho_3, k_i q_i\rho_2; m_i q_i\rho_1)\right] \\
& + (1-P_{it-1})R_{it}log\left[\Phi_3(k_i\gamma_2' z_{it-1}, m_i\psi' w_{it-1}, q_i\beta' x_{it-1}; k_i m_i\rho_3, k_i q_i\rho_2; m_i q_i\rho_1)\right] \\
& + (1-R_{it})log\left[\Phi_2(m_i\psi' w_{it-1}, q_i\beta' x_{it-1}; m_i q_i\rho_1)\right]
\end{aligned}
$$

where $k_i \equiv 2P_{it}-1$, $m_i \equiv 2R_{it-1}-1$, $q_i \equiv 2P_{it-1}-1$. The first term in the sample likelihood function corresponds to the contribution of an individual who was poor in the initial wave and was retained in the sample. The second term is the contribution of an individual who was non-poor in the initial wave and was retained in the sample. The third term is the contribution of an individual whose poverty status was observed in the initial wave, but who was not retained in the sample.

The presence of the trivariate standard normal distribution function in the sample log likelihood function means that estimation of the model is rather complicated, and because of this we rely on the simulation methods outlined in Gouriéroux and Monfort (1996), and presented in the context of an endogenous switching model in Cappellari and Jenkins (2006). Our estimation in this study uses the GHK simulator with 250 random draws.

One important part of the estimation process to consider is the fact that by pooling observations we are violating the maximum likelihood estimation assumption of independently and identically distributed observations. In order to correct for this we cluster our standard errors at the household level in the wave in which each respondent first appears in the data. For example, if a respondent appears in all four waves, the cluster is defined as the household identifier in wave 1. Respondents who are added to the sample from wave 2 onwards (such as TSMs) are allocated a cluster according to the household in which they are first observed. This allows us to account for arbitrary levels of intra-household correlation while maintaining the assumption of independence across households.

Identification in this model can come in two ways. One can either force the cross-equation correlations to equal zero, or invoke a set of exclusion restrictions that impact on initial poverty or the probability of attrition, but do not have an effect on the transitions themselves. The first method is not attractive, as one of the strengths of this particular way of thinking about poverty transitions is that we can, in fact, estimate these cross-equation correlations. This means we have to argue for identification of the model through a set of exclusion restrictions that allow us to test whether initial conditions and attrition are exogenous. If we find that $\rho_1 = \rho_2 = \rho_3 = 0$ then we can simply use univariate probit models to estimate poverty dynamics. Similarly, if we find that $\rho_1 = \rho_3 = 0$ then we can treat attrition as random and ignorable, and our estimation reduces to a bivariate probit model. Finally, if we find that $\rho_1 = \rho_2 = 0$, then we are able to treat poverty in $t-1$ as exogenous. These are interesting questions even if asked independently of the trivariate model itself.

The first set of exclusion restrictions requires us to find a variable (or a set of variables) that determine initial poverty status but are unrelated to the transition into or out of poverty. To this end we follow Cappellari and Jenkins (2004) and Jenkins (2011) who use the argument in Heckman (1981*b*) that initial conditions for labour market outcomes may be instrumented by information on the individual prior to labour market entry. In our model we use variables of the head of the household's parental occupation and education as instruments for initial conditions.[24] That is, these variables appear in x_{it-1} but not in z_{it-1}.

The second exclusion restriction requires us to include a variable (or set of variables) in w_{it-1} that is not included in z_{it-1}. That is to say we need a variable that affects the probability of attrition, but not poverty retention or transition. We take advantage of the panel structure of NIDS and include a dummy variable for whether a respondent is a continuing sample member (CSM) or a temporary sample member (TSM). CSMs are sample members who appear in the first wave of NIDS, while TSMs are those respondents who joined the household of a CSM in waves 2, 3 or 4. CSMs are tracked from wave to wave, while TSMs are not. In using this exclusion restriction we are asserting first that CSMs and TSMs have different probabilities of being retained in the sample, and second that the propensity to transition into or out of poverty

[24]NIDS includes a module in the adult questionnaire that records the education and occupational category of each respondent. The dummy variables include the following sectors: agriculture/elementary, professional, semi-skilled/operator, crafts, clerks, and a dummy variable for missing occupational sector.

is unrelated to whether a respondent is a CSM or a TSM.

The coefficient vectors in our model share a core of common variables, all of which are contained in z_{it-1}. The vector in the initial conditions equation, x_{it-1}, is comprised of z_{it-1} with additional variables controlling for parental education and occupation categories. Likewise, the vector in the sample retention equation, w_{it-1} contains z_{it-1} plus a dummy variable which identifies whether a respondent is a CSM or a TSM.

We are able to test the validity of our exclusion restrictions by treating the non-linear functional form of the model as being sufficient for identification, and using the parental background variables and CSM dummy as over-identifying restrictions for the initial conditions and sample retention equations, respectively. The results of these tests are presented in Table 3.9, along with our estimates of the ρ correlation coefficients.

The first panel of Table 3.9 shows the three ρ correlations that were defined earlier, all of which are statistically significant at the 1% level. The fact that ρ_1 is positive indicates that poor respondents in the initial period are less likely to have attritted compared to those who started off non-poor. This is unsurprising, as the disproportionate loss of wealthier households from wave to wave has been a feature of the panel dynamics of NIDS (de Villiers et al., 2013; Baigrie and Eyal, 2013). The correlation between unobservables affecting initial poverty and poverty transitions, ρ_2, is also positive, reflecting the fact that respondents who were poor in the initial period were more likely to be poor in the next period, compared to those who were non-poor to start off with. Finally the correlation between the unobservables determining the relationship between retention in the sample and conditional current poverty status, ρ_3 is negative. This implies that, for example, for the subsample of individuals who were poor in $t-1$, those who were retained in the sample are less likely to be poor in t than those who attritted would have been had they been retained.

Tests of the exogeneity of the different processes are contained in the second panel of Table 3.9. A Wald test of ρ_1 and ρ_2 is rejected at the 1% level, implying that initial conditions are not exogenous and should be accounted for in the modelling of poverty transitions. The null hypothesis of the exogeneity of sample retention is rejected at the 1% level (test statistic 59.73), indicating that sample retention is endogenous when modelling poverty transitions, and that accounting for it in the model is important. Finally, a joint test of ρ_1, ρ_2 and ρ_3 being zero is also rejected at the 1% level, confirming that initial conditions and sample retention are endogenous, and lending weight to our strategy of modelling poverty dynamics using a trivariate probit.

The next part of the table presents tests of the suitability of the instruments that are added to the vector of coefficients in the transition and retention equations. Wald tests show that parental background variables of the household head and sample membership status can be excluded from the transition equation both separately and jointly. In contrast, these were found to be statistically significant in the initial conditions equation and the retention equation, respectively. Thus it appears that the NIDS data support the use of these instruments in the estimation of poverty transitions and that we do not have to rely solely on non-linearity as our identifying factor. Finally, we tested for state dependence by calculating a Wald statistic for the equality of γ_1 and γ_2 from the equation estimating current poverty, and the null hypothesis of no state dependence was rejected at the 1% level.

Table 3.9: Model correlations and test statistics

Correlations between unobservables	**Estimate**	**p-value**
Initial conditions and retention (ρ_1)	0.039	0.000
Initial conditions and conditional current poverty (ρ_2)	0.196	0.000
Retention and conditional current poverty (ρ_3)	-0.228	0.000
Null hypotheses	**Test statistic**	**p-value**
Unobservables:		
Exogeneity of initial conditions (ρ_1=ρ_2=0)	56.06	0.000
Exogeneity of sample retention (ρ_1=ρ_3=0)	73.89	0.000
Joint exogeneity (ρ_1=ρ_2=ρ_3=0)	117.47	0.000
Transition equation:		
Exclusion of parental background (d.f.=10)	11.40	0.250
Exclusion of sample membership status (d.f.=2)	1.11	0.574
Exclusion of both (d.f.=12)	12.31	0.341
Initial conditions equation:		
Inclusion of parental background (d.f.=5)	42.10	0.000
Retention equation:		
Inclusion of sample membership status (d.f.=1)	3 439	0.000
State dependence:		
No state dependence, $\gamma_1 = \gamma_2$ (d.f.=50)	5 933	0.000

Source: Own calculations from the first four waves of NIDS.

The evidence presented suggests that our estimation strategy is sound. We turn now to the impacts of the independent variables on the probabilities of poverty transition and poverty persistence.

We calculate two sets of marginal effects - one set for poverty persistence and one for poverty entry, corresponding to poverty status in $t-1$. These are derived from the equations defining s_{it} and e_{it} respectively. We follow Stewart and Swaffield (1999) and Cappellari and Jenkins (2004) in defining the marginal effects of this model. The following explanation is related to the poverty persistence equation defining s_{it}. The corresponding explanation for poverty entry, e_{it} is constructed analogously. Because of the three related processes being modelled in this estimation strategy, a marginal change in one of the components of z_{it-1} will also result in a change in x_{it-1} because of the common elements in both vectors. This implies a change in the denominator of s_{it}, that is, the probability of being poor in $t-1$. In order to hold this constant in the calculation of the marginal effects we applied the following steps. We calculate the predicted probability of being poor in $t-1$ for all those respondents who were poor at this time. Next we take the average of these predicted probabilities which we call c. Then, into the denominator of the equation for s_{it} we substitute in $d \equiv \Phi^{-1}(c)$ which gives us the expression $\Phi_2(\gamma_1' z_{it-1}, d; \rho_2)/d$. Marginal effects for continuous variables are calculated by inducing an infinitesimal change in the covariate with all other covariates held constant at their means. For binary variables the marginal effect is calculated as the change in s_{it} implied by a unit change in the variable of interest, relative to a reference person. This reference person is defined by setting all continuous variables to their median values, and setting all binary covariates equal to zero.

The first two columns of results in Table 3.10 present the marginal effects and associated t-ratios for poverty in t conditional on being poor at $t-1$.[25] Females were four percentage points more likely to remain in poverty than males. African and coloured sample members had a conditional poverty probability that was around 31 percentage points higher than white sample members. Of the household head's characteristics, having a completed secondary or tertiary education was associated with a lower conditional poverty probability of 13.5 percentage points, relative to the base category of no education. The marginal effect of living in a household in which the head is employed is not statistically significant at the 5% level. Urban households were 6.4 percentage points less likely to remain in poverty than the base category of rural households. The presence of at least one employed household member is associated with a 5.8 percentage point reduction in the probability of remaining poor, while the presence of at least one child aged 15 or under results in a 4 percentage point increase in the probability of being poor in both time periods.

The third and fourth columns in the table present the marginal effects of conditional poverty entry between periods $t-1$ and t. The economic significance of each individual covariate is similar to the estimates of conditional poverty persistence, though the t-ratios of age, age squared and the female dummy variable are three to four times smaller. Africans are almost

[25]The table shows that there are 40 850 individuals who form part of the estimation sample, and three pairs of wave-to-wave transitions (wave 1 to wave 2, wave 2 to wave 3, and wave 3 to wave 4). The mapping to the number of person-waves is as follows. Balanced panel members appear three times in the wave-to-wave pairs, and each of these will show up in the number of person-waves considered. However, because we are using the pooled sample of NIDS respondents over all four waves, we also include those who experience one or two transitions (appearing in only two or three consecutive waves) in the estimation of the model.

30% more likely to enter poverty than the base group of whites, on average, even after controlling for individual, household head and overall household characteristics. The corresponding effect for coloured respondents is 23%. Living in a household with a female household head is associated with a higher conditional probability of poverty entry of 8.5 percentage points - more than four times larger than the effect for conditional poverty persistence. The protective effect of having at least a matric is more than twice as strong against the conditional probability of entering poverty than it is against the conditional probability of remaining in poverty. The largest difference in these columns compared to the previous two is that having at least one household member aged 65 and above increases the probability of poverty entry by almost 16 percentage points, whereas the effect was not statistically significant at the 5% level in the poverty persistence estimates.

Table 3.10: Model estimates of poverty in t, conditional on poverty status in $t-1$

Covariate at $t-1$	Poor at $t-1$		Non-poor at $t-1$	
	Marginal effect	|t-ratio|	Marginal effect	|t-ratio|
Individual				
Age	-0.002	(11.37)	-0.003	(3.62)
Age squared	0.00004	(8.62)	0.00004	(2.12)
Female	0.040	(8.82)	0.042	(2.46)
African	0.314	(11.71)	0.294	(13.71)
Coloured	0.312	(11.54)	0.231	(10.77)
Household head				
Age	-0.001	(2.05)	-0.005	(7.11)
Age squared	0.00002	(3.03)	0.00002	(4.92)
Female	0.021	(4.35)	0.085	(4.54)
Matric and above	-0.135	(17.18)	-0.279	(17.85)
Employed	0.009	(1.53)	-0.039	(1.84)
Household				
Urban	-0.064	(10.69)	-0.084	(3.77)
Farm	-0.005	(0.46)	-0.029	(0.78)
Adult 65 and above	0.009	(1.24)	0.157	(4.74)
Children 15 and below	0.040	(25.67)	0.028	(3.12)
Any workers	-0.058	(9.88)	-0.006	(0.20)
Own dwelling	0.041	(6.56)	0.006	(0.29)
Log likelihood	-94 355			
Model chi-squared (d.f. = 54)	4 306 ($p<0.000$)			
No. of clusters	13 238			
No. of observations	40 850			
No. of observations (person-waves)	88 090			

Source: Own calculations from the first four waves of NIDS. Reference categories for binary covariates: male, white, male household head, household head does not have matric, household head is not employed, Western Cape province, rural area, no adults over 65 in the household, no children under 15 in the household, no workers in the household, household members do not own the dwelling. The base wave is wave 1.

Another way of interrogating the findings is to run the full Markovian model on the subsample of African respondents only. As was shown in the poverty transition matrices earlier in the chapter, although the African subsample drives most of the results, there are often important differences in the dynamics between African and non-African panel members. Table 3.E.2 in the appendix estimates the marginal effects from the same type of Markovian model that was reported in Table 3.10, but restricts the analysis to African respondents only.[26] The change in

[26]Full results along with the estimated correlations between the unobservables are available from the author.

sample also implies a change in the underlying correlations between unobservables, and the underlying distributions of the explanatory variables. The base category for the calculation of the marginal effects remains the same except for the fact that there is no longer a base category for race, as the variable does not vary in this subsample.

The marginal effects for poverty persistence are relatively similar when comparing the full sample to the African subsample. However, the protective effect of a household head having at least a matric education decreases from 13.5% to 11.5%. The marginal effect on poverty persistence of living in an urban area, relative to the rural base category, decreases in absolute terms from -6.4% to -5.5%, on average.

There are more differences between the samples in the poverty entry marginal effects, compared to the poverty persistence marginal effects. The individual marginal effects of age and gender for the poverty entry model are very similar for the overall sample and for the African subsample. In the full sample having a female household head is associated with an 8.5% increased probability of transitioning into poverty, on average. This decreases to 6% when the model is estimated on the African subsample only. The protective effect of living in a household in which the head has at least a matric is lower in the African subsample – 23.1% compared to 27.9% in the full sample. Living in a household in which the head is employed has a larger effect amongst Africans than non-Africans in protecting against poverty entry, and this is in contrast to the smaller protective effect of living in urban rather than rural areas for this group. Interestingly, for the African subsample there is no statistically significant effect of the presence of an elderly person in the household on the probability of transitioning into poverty, in contrast to the overall effect (and therefore the effect amongst non-Africans). As was the case previously, the presence of children in the household has no statistically significant effect on the probability of transitioning into poverty, on average.

Although contexts and methodologies vary greatly between countries, it is worth pointing out some of the other existing research on poverty transitions that uses a similar estimation strategy to the one found in this chapter. The seminal Cappellari and Jenkins (2004) study uses BHPS data for Britain. The authors find statistically significant correlations between initial conditions and retention (ρ_1), and between initial conditions and current poverty (ρ_2). Although there is no statistically significant relationship between the unobservables affecting retention and the unobservables affecting conditional current poverty (ρ_3), the full set of unobservable correlations are jointly significant. The results in this study suggest that living in a household in which the head has at least completed A-levels is associated with a lower probability of poverty persistence. The presence of children in the household significantly increases the probability of conditional poverty persistence. In general, there are more statistically significant relationships in the poverty entry equation, compared to the poverty persistence equation. Older respondents and households with male heads are both associated with higher probabilities of transitioning into poverty, as are single parent families and multi family households. Another example of a Markovian approach to studying poverty transitions in an OECD country is Buddelmeyer and Verick (2008) which uses the first five waves of the HILDA longitudinal dataset in Australia. This paper finds that poverty is largely a transient phenomenon in Australia, and that having a tertiary education is a large buffer against both poverty persistence and poverty entry. The study also uncovers an important geographic aspect to who becomes who and who remains in poverty, with poverty concentrated in remote and rural areas of the country.

Although there are few examples of this data-intensive approach to studying poverty dynamics in developing countries, Faye et al. (2011) use the 3rd and 13th waves of a large dataset which tracks welfare in Nairobi slums. Their results suggest that in this context, only the correlation between the unobservables affecting initial conditions and the unobservables affecting retention are statistically significant. The presence of children in the household is associated with a higher probability of remaining in poverty, but not of entering poverty. Neither the gender of the respondent nor that of the household head has a significant effect on either poverty persistence or poverty entry. Finally, Azomahou and Yitbarek (2015) follow a sample of 837 households in Ethiopia in order to study poverty transitions. The results of this study show that the education level of the household head is an important buffer against poverty persistence, but that it has no significant effect on the probability of poverty entry. Perhaps surprisingly, there are once again no gender differentials in the probability of remaining in or transitioning into poverty.

How does estimating conditional poverty dynamics in this way change our understanding compared to a simple univariate probit model? Table 3.F.1 in the appendix presents results from a univariate probit model using the same vector of covariates that as in Table 3.10, except without controlling for initial conditions and non-random attrition. The marginal effects in the Markovian model are generally more economically significant than they are in the probit. For example, being female increases the conditional poverty entry probability by 4.2 percentage points in the Markovian model, but only by 1.4 percentage points in the univariate probit. The presence of at least one adult aged 65 and above increases conditional poverty entry in the Markovian model by almost 16 percentage points, while the corresponding increase in the univariate probit stands at 5 percentage points. The protective value of having at least a matric is heightened in the Markovian estimation with a marginal effect of -28 percentage points, compared to almost -16 percentage points in the univariate probit. One interesting point to observe is that the sign of the dummy variable denoting the presence of children aged 15 and below in the household switches from negative to positive as we move from the probit model to the Markovian estimation in the poverty entry equations.

What do the results suggest about the length of poverty spells?

In an early paper using a first-order Markovian model to estimate transitions into and out of welfare, Boskin and Nold (1975) show that the conditional probabilities of being in each state follow a geometric distribution and can be used to generate statistics on the length of time that each sample member can expect to be in a given state. In the South African context, Carter and May (2001) also assume a stationary Markov process in order to try to uncover the long run distribution of poverty status using six different categories of welfare.[27] For the trivariate case, Jenkins (2011) shows that, assuming stationarity, equations 3.9 and 3.10 can be used to calculate the average length of time that an individual is expected to be in poverty. This is given by $1/(1 - s_i)$, while the median duration is given by $log(0.5)/log(s_i)$.[28] The average length of time that an individual will spend out of poverty is $1/(e_i)$, with median time out of poverty given by $log(0.5)/log(1 - e_i)$. Finally, the unconditional probability that an individual is poor

[27]This is a slightly different application to what we are interested in, as the authors uncover the long-run distributions of poverty classes, while we are more interested in relative poverty and non-poverty spell lengths.

[28]The time subscript has been omitted because of the assumption of stationarity.

is expressed as $e_i/(e_i + 1 - s_i)$. Given the way in which the model was estimated, these spell length estimates control for the biases introduced by initial conditions and non-random attrition by construction.

As noted in the introduction to the Markovian model, projections like these rely on the assumption of no state dependence (thus allowing for left-censored poverty spells to be incorporated). In practice this means that predictions of mean and median poverty spell lengths rely on a fundamental stationarity assumption. The choice of base category and subsequent variations in Table 3.11 therefore relies on variables that are more likely to be stable over time (for example education of the household head, race and gender). In addition, rather than displaying the predicted lengths of poverty and non-poverty spells, the table and discussion present the lengths of these spells relative to the base category.

Table 3.11 presents the predicted poverty transition probabilities, steady state probabilities and relative spell lengths of poverty and non-poverty for a range of different characteristics. In the first case, the reference person is a 40 year old African male living in urban KwaZulu-Natal in a household in which the household head has at least a matric level of education, in which there are no children under the age of 15 and in which there are no adults aged 65 and over. The poverty persistence rate associated with this individual is 0.435, while the predicted poverty entry rate is 0.229. The predicted probability of this man being poor is just under 0.3. Given the stationarity assumption, the average lengths of time spent poor and non-poor have both been set to 1, so that all other results can be interpreted relative to this base category. In practice the mean length of time spent poor or non-poor is higher than the median for the base category and all subsequent comparisons. This reflects the relatively wide distribution and relatively long right hand tail of poverty spell lengths for individuals with the same characteristics.

In the next row of Table 3.11 we change the sex of the reference person to female, while keeping all other characteristics the same. This increases the predicted probability of remaining in poverty by about four percentage points, and the predicted probability of entering poverty by almost three percentage points. The overall probability of being poor is 33%, which is four percentage points higher than a male with the same characteristics. This number is broadly in line with the marginal effects presented in Table 3.10, though the reference characteristics are different. The implication of these relatively higher persistence and entry rates is that the mean and median poverty spell lengths are longer, while the mean and median non-poverty spell lengths are shorter, relative to the base category.

In the third row we keep the characteristics of the person in row 1, except for changing the racial group from African to white. In line with the results of the model and the descriptive statistics, this has a very large impact on the predicted poverty states. The predicted probability of remaining poor is just under 13%, compared to 43.5% for African respondents with the same characteristics, while the predicted probability of entering poverty is just 3%, compared to 23% for African respondents with the same characteristics. The average length of time that a person with these characteristics can expect to remain poor 35% lower for whites than for Africans, and the mean length of time spent non-poor is 7.67 times longer for whites than for Africans.

The fourth row keeps the same characteristics as the first row apart from changing the education of the household head from matric and above to less than matric. In line with the large economic significance reported in Table 3.10, this single change increases the probability of an individual being in poverty from 29% to 60%. The predicted poverty persistence and poverty entry rates are higher, as are the mean and median lengths of time for the predicted spell length of poverty, relative to the base category.

The remaining rows show that the highest probability of being poor is for individuals living in a household with children aged 15 and under, with at least one adult aged 65 and above, and with a household head who does not have a matric level of education. This predicted probability is 70%, over 40 percentage points higher than the base category in row 1. The longest relative average poverty spell is for those individuals who live in a rural household with a household head who does not have a matric level of education (see row (8)). As can be seen in row (10), an individual who lives in a household with children under 15 years old and elderly adults, in which the household head does not have a matric can expect to be in poverty for 46% longer than the base category, on average, and can expect to experience spells of non-poverty that are only about a quarter as long as those experienced by the base category, on average. The median poverty spell length for someone with these characteristics is 71% longer than that experienced by the base category, while the median length of time spent non-poor is only 12% of the time that someone with the base category characteristics can expect to experience.

Table 3.11: Predicted transition probabilities, steady-state probabilities and relative spell lengths

Characteristics	Poverty persistence rate (s_{it})	Poverty entry rate (e_{it})	Pr(poor)	Poverty spell length relative to (1)		Non-poverty spell length relative to (1)	
				Mean	Median	Mean	Median
(1) Male, aged 40, African, male HHH with matric, KZN province, urban, no adults 65 and over, no children 15 and under	0.435	0.229	0.289	1.00	1.00	1.00	1.00
(2) As (1) except female	0.474	0.257	0.328	1.07	1.12	0.89	0.88
(3) As (1) except white	0.128	0.030	0.033	0.65	0.41	7.67	8.58
(4) As (1) except HHH no matric	0.573	0.641	0.601	1.33	1.50	0.36	0.25
(5) As (4) except female	0.606	0.698	0.639	1.44	1.66	0.33	0.22
(6) As (4) except female HHH	0.622	0.819	0.684	1.50	1.76	0.28	0.15
(7) As (4) except with adults over 65	0.581	0.851	0.670	1.35	1.53	0.27	0.14
(8) As (4) except rural	0.627	0.775	0.675	1.51	1.78	0.30	0.17
(9) As (7) except workers in HH	0.528	0.841	0.641	1.20	1.31	0.27	0.14
(10) As (7) except with children in the HH	0.614	0.896	0.699	1.46	1.71	0.26	0.12

Source: Own calculations from the first four waves of NIDS. HH = household. HHH = household head.

How sensitive are our results to the choice of poverty line?

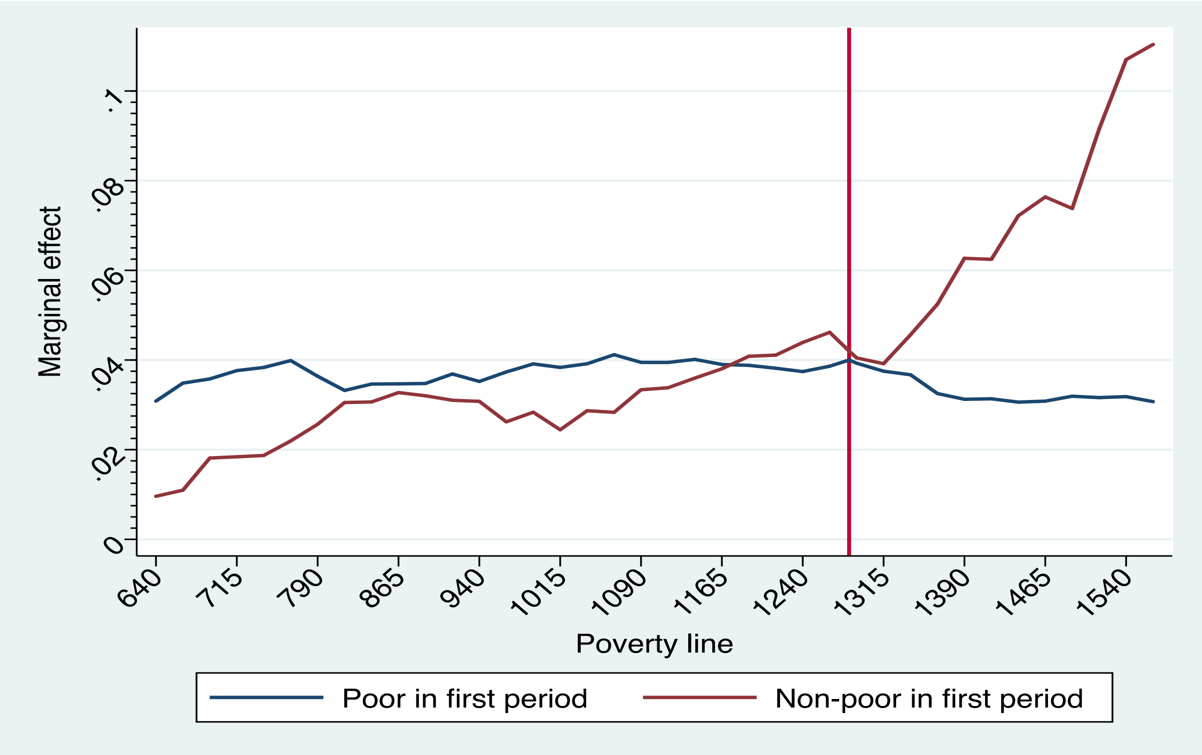
The choice of which poverty line to use always involves some level of arbitrariness, and one obvious robustness check is to investigate how results change for different poverty lines. Using poverty status as a binary dependent variable is very different to using other binary dependent variables in estimation. For example, it is generally objective whether an individual is employed or not, whether an individual receives a pension or not, or whether a child is enrolled in school or not. Converting money metric welfare into a binary variable reflecting poverty, however, necessarily involves some subjective judgements. This means that checking the sensitivity of the model's results for different poverty lines is important. The primary reason for this is so that we are not 'hostage to an…arbitrarily selected poverty line' Deaton (1997). A second reason for doing so is that it may be interesting in itself to see how the effect of a particular variable on poverty transitions changes as the definition of poverty changes. For example, we may be interested in whether the effect of having a household head with a matric education has an increasingly protective effect against falling into poverty as the poverty line rises.

We use a wide range of possible poverty lines in order to see both how marginal effects change from a low to a high poverty line, and also to see how stable the marginal effects are in the neighbourhood of our poverty line of R1 283. The lower bound for our poverty line range is R640, which is just under half the poverty line used in the estimation. It is also close to the StatsSA lower bound poverty line of R608 (Statistics South Africa, 2015), and it is unlikely that any reasonable poverty line for South Africa would be below this amount. The upper bound for our robustness check is a very high poverty line of R1 565, giving us a total window of just over R900 in which to assess the sensitivity of our results. The poverty line of R1 283 is the highest line in general use in the South African academic discourse, and the results for poverty lines above this level should be interpreted with this in mind

In Figure 3.5 we show the marginal effects for the female dummy variable for poverty lines from R640 to R1 540. Recall from the table of results that females who were poor in period $t-1$ were, on average, 4 percentage points more likely to remain in poverty than males. The corresponding marginal effect for females who were not in poverty in $t-1$ was almost the same, at 4.2 percentage points. This small difference is reflected at the vertical line in the figure, corresponding to the model's poverty line of R1 283.

The female marginal effect for those who were poor in $t-1$ is far more stable than the corresponding marginal effect for females who were non-poor in $t-1$. In fact, any poverty line from R950 to R1 350 would give a marginal effect of around 4 percentage points for females who remain in poverty. The marginal effect for females entering poverty is relatively stable in the vicinity of the R1 283 poverty line, but it explodes thereafter - jumping by over 150% from a poverty line of R1 283 to a poverty line of R1 540. Had we chosen the StatsSA lower bound poverty line, we would have seen female marginal effects of 3 percentage points and 1 percentage point for poverty retention and poverty entry, respectively.

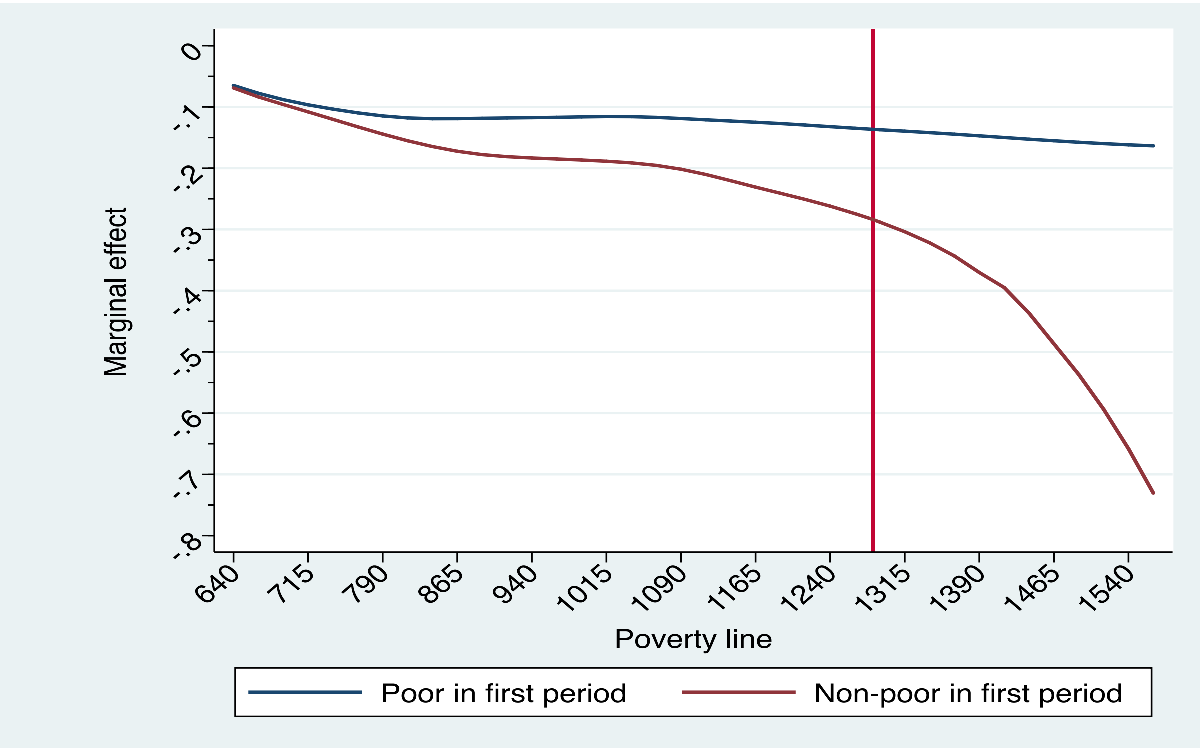
Figure 3.5: Female marginal effect on the probability of being poor in the second period for different poverty lines



Source: Own calculations from the first four waves of NIDS.

The sensitivity of the marginal effect of having a household head with at least a matric is interesting because it is the single largest marginal effect apart from the race categories (in absolute terms) in both results columns of Table 3.10. While the female marginal effects diverge for poverty lines higher than R1 283, the marginal effects of having a household head with at least a matric begin to diverge much earlier. This can be seen in Figure 3.6, where the marginal effects at a poverty line of R640 are identical at -6 percentage points. Had we chosen this lower bound poverty line, we would not have been able to identify any discernible difference between the two different states in $t-1$. The 'protective' effect against remaining in poverty of having a household head with at least a matric increases steadily up to our poverty line of R1 283, where it stands at -13.5 percentage points. From there the effect increases gradually in absolute terms to -16.5 percentage points for a poverty line of R1 540. The lower line in Figure 3.6 shows very different marginal effects across the distribution of poverty lines. These significant differences over the range of poverty lines are likely driven by the fact that relatively few people living in a household in which the head of the household has at least a matric actually fell into poverty between $t-1$ and t. At our poverty line of R1 283 we find that the economic significance of having a household head with at least a matric is double for those who were non-poor in the first period, compared to those who were poor, on average. This difference would be zero if we used the lowest poverty line in the figure, but would be almost of the order of 4.5 if we used a poverty line as high as R1 540.

Figure 3.6: Household head with matric or above marginal effect on the probability of being poor in the second period for different poverty lines



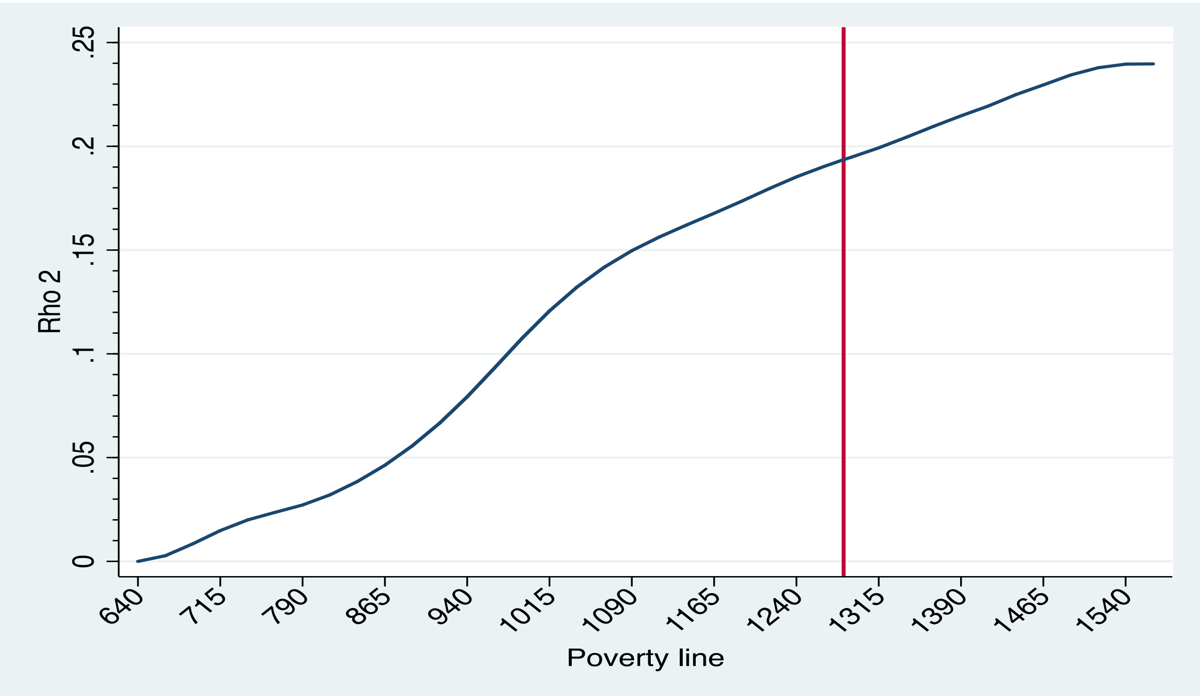
Source: Own calculations from the first four waves of NIDS.

We also check the sensitivity of the results if we had ignored initial conditions and non-random attrition, and instead estimated a probit model of poverty transitions. The marginal effects for the female and household head with at least a matric covariates for this can be seen in figures 3.G.1 and 3.G.2 in the appendix. In general the marginal effects from the probit models do not change as much for higher poverty lines as they do in our Markovian model. The marginal effect for females who were both poor and non-poor in $t-1$ generally decreases as the poverty line increases. Interestingly, while the Markovian model produced diverging effects, conditional on initial poverty status, for those living in a household in which the head had at least a matric, the univariate probit model shows convergence over the same range of poverty lines, to the point where the difference is negligible for the highest poverty line of R1 540.

Figures 3.G.3 and 3.G.4 compare the marginal effects of the African and coloured variables for the poverty persistence versions of the Markovian and probit models respectively. In both cases the marginal effects from the Markovian model are higher than the marginal effects of the probit model, at any reasonable poverty line. The difference between the marginal effects is highest at the lowest poverty line of R640 - in both cases the Markovian marginal effect is about double the probit marginal effect. The marginal effects from both models converge as the poverty line increases, and are quite close to one another at the highest poverty line of R1 565. As shown in Table 3.10 and Table 3.F.1, the African and coloured marginal effects are extremely high, relative to the base category of white. On average, Africans are 31% and 24% more likely to remain in poverty than whites according to the Markovian model and the probit model respectively. The Markovian model shows that the marginal effect for coloured respondents remaining in poverty is also 31%, relative to the base category, while the corresponding marginal effect from the probit model is that this group is 23% more likely to remain in poverty than whites, on average. Interestingly, although the African marginal effect is generally slightly below the coloured marginal effect, the shape of the line of marginal effects over the range of poverty lines is similar for both groups, whether a Markovian model or a probit model is used in the estimation.

One of the reasons for the large differences in the marginal effects at different poverty lines between the Markovian and univariate probit estimations is the change in ρ_2 over the range of poverty lines. Recall that ρ_2 enters the calculation of the marginal effects via its presence in the numerator of equation 3.9 and equation 3.10. That is to say, the correlation between the unobservables affecting initial conditions and conditional current poverty changes over the range of poverty lines. In fact, as shown in Figure 3.7, the estimated correlation is very close to zero for the lowest poverty line, and then rises steadily until it reaches 0.196 at the poverty line of R1 283. The growth in the correlation over the range of poverty lines shows that state dependence becomes increasingly important as the poverty line increases, and reinforces the fact that our findings cannot remain agnostic to the choice of poverty line. The intuition behind the positive slope shown in this figure is that a) a lower poverty line means fewer people will be found to be in poverty and b) there will be more transitions out of poverty compared to a higher poverty line, by construction. More transitions out of poverty means a lower level of state dependence, which implies a lower correlation between the initial conditions and the conditional current poverty status. Therefore the positive slope of the line is a result of a higher poverty line being associated with a higher correlation between initial conditions and conditional current poverty.

Figure 3.7: Correlation between unobservables affecting initial conditions and conditional current poverty status for a range of poverty lines



Source: Own calculations from the first four waves of NIDS.

What do the results tell us about aggregate state dependence versus genuine state dependence?

An attractive feature of modelling poverty transitions in the way that we have is that it allows us to distinguish between aggregate state dependence (ASD) and genuine state dependence (GSD). ASD is simply the unconditional difference between the probability of being poor in

t for those who were poor in $t-1$, and the probability of being poor in t for those who were non-poor in $t-1$. Measuring state dependence in this way does not take account of individual heterogeneity. It is easily calculated using the top panel of Table 3.8 as follows:[29]

$$ASD = \left(\frac{\sum\limits_{i\epsilon\{P_{it-1}=1\}} Pr(P_{it}=1|P_{it-1}=1)}{\sum\limits_{i} P_{it-1}} \right) - \left(\frac{\sum\limits_{i\epsilon\{P_{it-1}=0\}} Pr(P_{it}=1|P_{it-1}=0)}{\sum\limits_{i} (1-P_{it-1})} \right)$$

The calculation of GSD controls for observed and unobserved individual-level heterogeneity, and is particularly important if initial conditions matter for poverty in the current period. As explained in Cappellari and Jenkins (2004), "the experience of poverty itself might induce a loss of motivation, lowering the chances that individuals with given attributes escape poverty in the future." One possible formal test for GSD is to test the null hypothesis that γ_1 and γ_2 from equation 3.5 are equal. The result of this test is shown in the final row of Table 3.9, in which the null hypothesis of no GSD is decisively rejected. Our equation for calculating GSD involves calculating the difference between the predicted probability of being poor in t, if poor in $t-1$, and the predicted probability of being non-poor in t, if non-poor in $t-1$ using the equations defining s_{it} and e_{it}. This is then summed over all individuals and divided by the number of individuals in the sample as follows:

$$GSD = \left(\frac{1}{N}\right) \sum_{i=1}^{N} \left[Pr(P_{it}=1|P_{it-1}=1) - Pr(P_{it}=1|P_{it-1}=0) \right]$$

The distinction between ASD and GSD is useful because the policy implications depend on the relative importance of each kind of dependence. As argued in Arulampalam et al. (2000), "Identification of the extent of true state dependence...is more than just an academic exercise."[30] If there is very little GSD of poverty at the individual level, then short-run interventions to reduce poverty will have little impact, as poverty will mainly be generated by adverse individual heterogeneity. However, if there is a high level of GSD, then reducing the probability of the

[29] In the ASD equation $\sum\limits_{i} P_{it-1}$ is the number of individuals who were poor in $t-1$, as for these individuals $P_{it-1}=1$. Likewise, $\sum\limits_{i}(1-P_{it-1})$

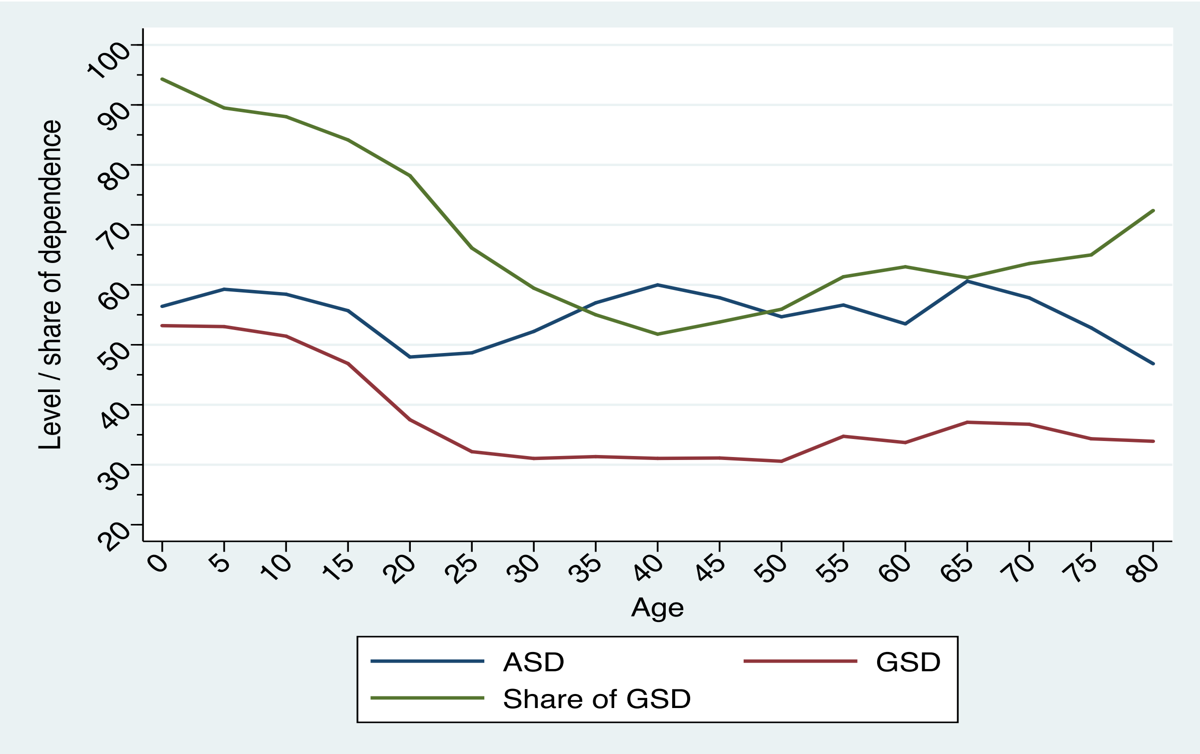
[30] Arulampalam et al. (2000) study unemployment persistence rather than poverty persistence, but their insights generalise to our context.

initial experience of poverty becomes crucial in the process of reducing long-run poverty in the country. This echoes Devicienti and Poggi (2011) who argue that the greater the share of GSD in overall state dependence, the higher the payoffs are to short-term income support programmes that prevent people from falling into poverty today. If, however, poverty persists mainly because of individual heterogeneity, then short-term income support schemes will have little effect on the long-run distribution of poverty in the country.

Our model estimates a level of ASD of 0.568 and a corresponding level of GSD of 0.416. This implies that the share of genuine state dependence in overall state dependence is very high at 73.30%. Figure 3.8 presents three statistics of interest, ASD, GSD and the share of GSD in total dependence across the age range of sample members.[31] ASD is relatively stable and is generally between 50 and 60, while GSD stands at about 53 for the youngest age cohort and then drops to around 30 for those between the ages of 25 and 50. The shape of the green line representing the share of GSD in overall dependence falls at first, in line with the decline in GSD, and then rises for those aged 40 and above. Even at its lowest point, genuine state dependence accounts for more than half of total poverty dependence in South Africa.

[31] Dependence levels were calculated for five-year age bands from 0 to 80. Those older than 80 were excluded because of low observation numbers.

Figure 3.8: ASD, GSD and the share of GSD by age group



Source: Own calculations from the first four waves of NIDS. Note: observation numbers for age cohorts are as follows. 0-4 (8 193), 5-9 (8 609), 10-14 (8 706), 15-19 (8 268), 20-24 (5 136), 25-29 (4 037), 30-34 (3 861), 35-39 (3 512), 40-44 (3 372), 45-49 (3 019), 50-54 (2 612), 55-59 (1 979), 60-64 (1 486), 65-69 (1 133), 70-74 (752), 75-80 (450).

3.7 Conclusion

In this chapter we began our investigation into the dynamics of poverty in South Africa by using the balanced four wave sample of NIDS comprising 17 265 respondents to analyse poverty dynamics in South Africa from 2008 to 2014/2015. Using a poverty line of R1 283 in January 2015 rands we found that the rate of exiting poverty was higher between waves 2 and 3, and between waves 3 and 4 than it was between waves 1 and 2. About 47% of the sample was below the poverty line in each of the four waves in which they were interviewed. Transition matrices showed that 54% of the balanced panel were poor in both wave 1 and wave 4, with more than half in 'severe' poverty - defined as having real household income per capita of less than half the poverty line.

The importance of demographic events in shaping dynamics was highlighted by the role of household composition changes as drivers of poverty entry and exit. Inter-wave demographic changes were the main triggers for 56% of those who entered poverty and 59% of those who exited poverty between wave 1 and wave 4. One needs panel data such as NIDS to disentangle these demographic events from income events. The value of such work is shown here in highlighting the central importance of migration and household instability in driving who gets ahead and who falls behind in contemporary South Africa.

The increasing longer-run importance of access to government grants was highlighted, with grant income being the main trigger precipitating poverty exit for 23% of previously poor balanced panel members between wave 1 and wave 4. By implication this flags as a major concern the muted role of the labour market in driving the dynamics of poverty exit between 2008 and 2014/2015.

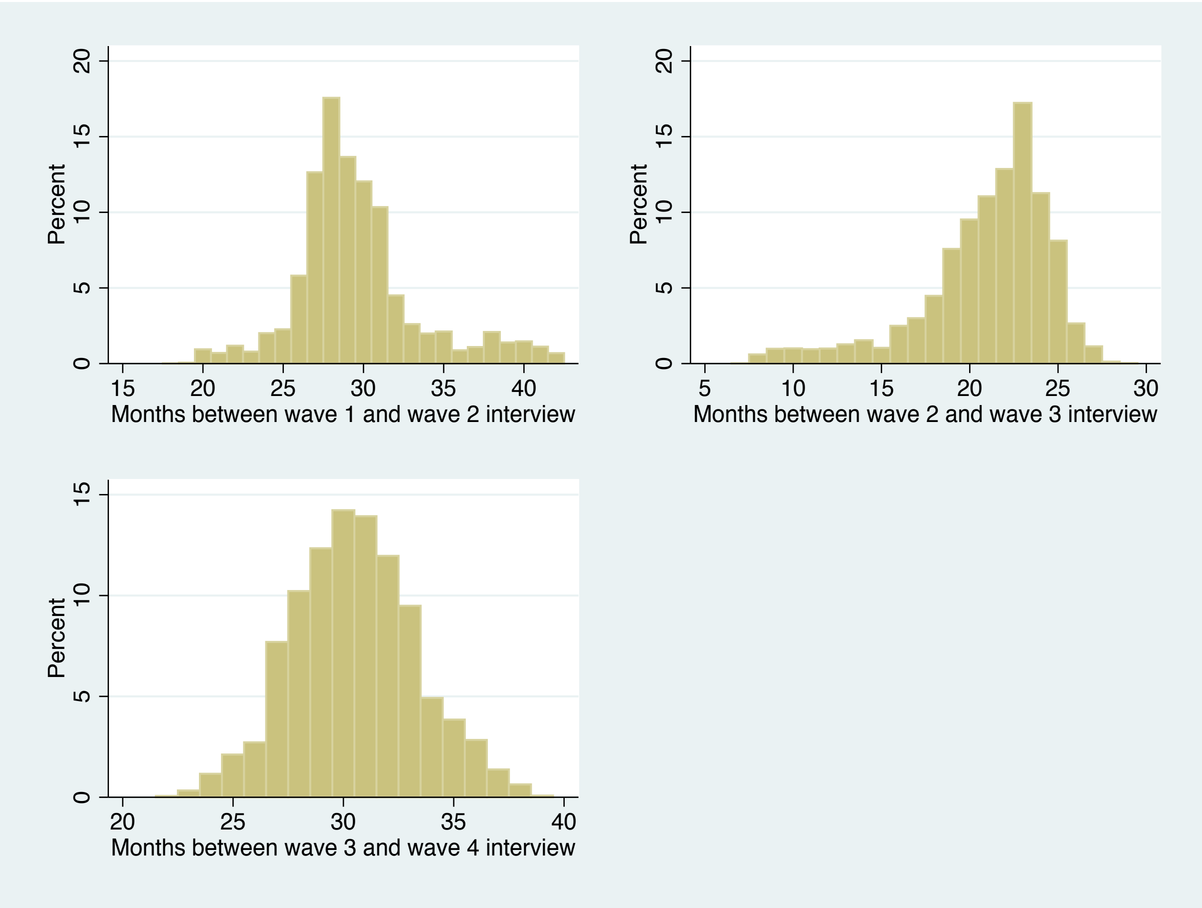
We then turned our attention to modelling poverty dynamics using a Markovian approach that simultaneously estimated poverty transitions along with initial conditions and selective attrition. We found that if researchers ignore the correlations between unobservables affecting initial conditions, sample retention and poverty transitions, then this could lead to substantially biased results. We also separated state dependence into the part attributable to aggregate state dependence and the part attributable to genuine state dependence, and found that the latter is dominant. From a policy perspective, this implies that preventing people from falling into poverty in the first place will likely yield greater returns in the long-run, rather than targeting the individual correlates of poverty directly.

Taken together, our results add to the body of literature showing that even after 22 years of democracy in South Africa, a very large proportion of its people have been unable to realise the economic freedom that should have come with political freedom. The task of us as researchers and policymakers is to ensure that this void is bridged as swiftly and justly as possible.

Appendix

3.A Number of months between being interviewed

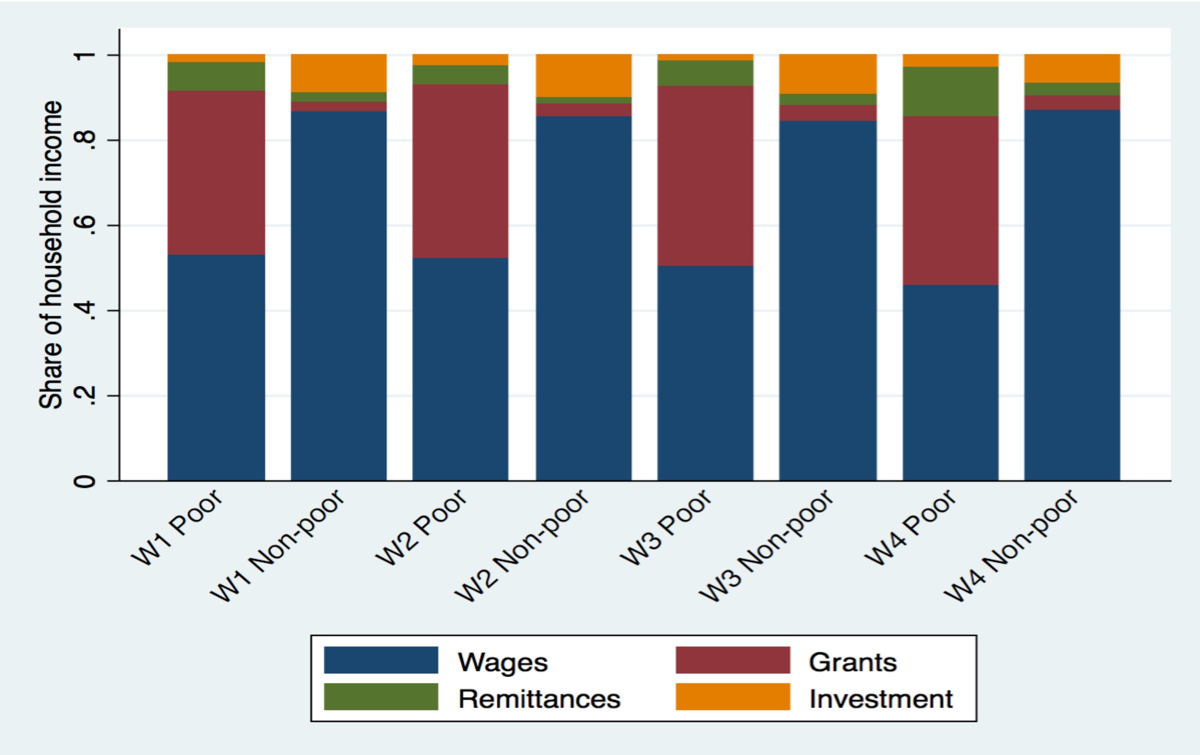
Figure 3.A.1: Number of months between interviews by wave interval for balanced panel members



Source: Own calculations from the first four waves of NIDS.

3.B Composition of household income

Figure 3.B.1: Household income composition for poor and non-poor respondents by wave



Source: Own calculations from the first four waves of NIDS. Attrition-corrected panel weights applied to members of the balanced panel.

3.C Transition matrices for African respondents

Table 3.C.1: Transitions into and out of poverty across waves: African balanced panel members only

		Wave 2				**Wave 3**	
		Poor	**Non-poor**			**Poor**	**Non-poor**
Wave 1	**Poor**	89.55	10.45	**Wave 2**	**Poor**	85.56	14.44
	Non-poor	35.59	64.41		**Non-poor**	26.47	73.53
		Wave 4				**Wave 4**	
		Poor	**Non-poor**			**Poor**	**Non-poor**
Wave 3	**Poor**	80.41	19.59	**Wave 1**	**Poor**	74.65	25.53
	Non-poor	25.77	74.23		**Non-poor**	28.48	71.52

Source: Own calculations from the first four waves of NIDS. Attrition-corrected panel weights applied to members of the balanced panel.

Table 3.C.2: Poverty transitions: Proportion of sample by transition status: African balanced panel members only

		Wave 2				**Wave 3**	
		Poor	**Non-poor**			**Poor**	**Non-poor**
Wave 1	**Poor**	72.41	8.45	**Wave 2**	**Poor**	67.79	11.44
	Non-poor	6.81	12.33		**Non-poor**	5.50	15.28
		Wave 4				**Wave 4**	
		Poor	**Non-poor**			**Poor**	**Non-poor**
Wave 3	**Poor**	58.93	14.36	**Wave 1**	**Poor**	60.36	20.50
	Non-poor	6.88	19.83		**Non-poor**	5.45	13.69

Source: Own calculations from the first four waves of NIDS. Attrition-corrected panel weights applied to members of the balanced panel.

Table 3.C.3: Transitions with finer poverty levels: African balanced panel members only

		Wave 2		
		Severe	**Poor**	**Non-poor**
	Severe	74.44	18.75	6.81
Wave 1	**Poor**	45.23	33.62	21.15
	Non-poor	17.95	17.64	64.41

		Wave 3		
		Severe	**Poor**	**Non-poor**
	Severe	64.69	24.82	10.49
Wave 2	**Poor**	38.42	36.61	24.98
	Non-poor	10.64	15.84	73.53

		Wave 4		
		Severe	**Poor**	**Non-poor**
	Severe	60.92	24.31	14.77
Wave 3	**Poor**	35.71	35.67	28.62
	Non-poor	12.13	13.64	74.23

		Wave 4		
		Severe	**Poor**	**Non-poor**
	Severe	54.67	24.29	21.05
Wave 1	**Poor**	30.23	31.79	37.98
	Non-poor	11.88	16.60	71.52

Note: In this panel the cells sum to 100%

		Wave 4		
		Severe	**Poor**	**Non-poor**
	Severe	32.97	14.65	12.69
Wave 1	**Poor**	6.21	6.53	7.80
	Non-poor	2.27	3.18	13.69

Source: Own calculations from the first four waves of NIDS. Attrition-corrected panel weights applied to members of the balanced panel.

3.D Trigger events with a modified equivalence scale

Table 3.D.1: Trigger events associated with poverty entry and exit: Modified equivalence scale

	Poverty entry				Poverty exit			
	W1 to W2	**W2 to W3**	**W3 to W4**	**W1 to W4**	**W1 to W2**	**W2 to W3**	**W3 to W4**	**W1 to W4**
Demographic								
Head changed	36.05	50.79	51.87	47.88	35.96	49.70	48.77	53.74
Needs > money	9.70	3.74	10.99	11.04	0.81	1.26	0.79	0.52
Demographic share	**45.75**	**54.53**	**62.86**	**58.92**	**36.77**	**50.96**	**49.56**	**54.26**
Income								
Head labour earnings	24.64	12.14	9.38	16.79	13.49	16.99	5.28	4.52
Spouse labour earnings	2.05	5.55	2.07	1.65	2.03	2.48	3.00	1.68
Other labour earnings	11.83	12.18	11.95	6.85	24.10	14.40	12.48	12.44
Remittances	3.95	3.10	3.28	3.88	2.43	3.87	7.56	3.56
Grant income	5.65	3.31	3.09	7.38	8.85	8.74	20.06	22.71
Income share	**48.12**	**35.85**	**29.77**	**36.56**	**56.21**	**42.98**	**48.39**	**44.92**
Inconclusive	6.13	9.62	7.37	4.52	5.99	6.06	2.05	0.62
Total	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
Observations	926	1 233	1 342	479	2 122	2 855	3 990	5 945

Source: Own calculations from the first four waves of NIDS. Attrition-corrected panel weights applied to members of the balanced panel.

3.E Descriptive statistics for sample included in Markovian model

Table 3.E.1: Descriptive statistics for sample included in Markovian model

	Poor in t-1	Non-poor in t-1
Individual		
Age	23.42	30.59
Race		
African	87.22%	59.71%
Coloured	11.94%	19.94%
Asian/Indian	0.44%	3.96%
White	0.39%	16.38%
Gender		
Male	44.79%	50.31%
Female	55.21%	49.69%
Retention		
Retained to period t	83.49%	76.60%
Household head		
Age	51.34	46.19
Female	60.67%	35.84%
Matric and above	8.49%	47.94%
Employed	34.75%	71.37%
Household		
Province		
W. Cape	9.04%	22.09%
E. Cape	14.06%	7.63%
N. Cape	6.82%	9.06%
Free State	5.43%	5.38%
KZN	33.27%	15.49%
North West	6.47%	6.31%
Gauteng	7.12%	20.30%
Mpumalanga	7.36%	7.95%
Limpopo	10.42%	5.80%
Geo-type		
Traditional	53.26%	17.01%
Urban	38.78%	75.43%
Farming	7.64%	7.42%
Composition		
Adults 65 and above	27.18%	15.82%
Children 15 and below	88.77%	61.76%
Any workers	57.90%	91.53%
Ownership		
Own dwelling	83.35%	71.21%
Observation numbers		
Individuals	40 850	
Person-waves	88 090	

Source: Own calculations from the first four waves of NIDS.
Variables reported at the individual level in period $t-1$.

Table 3.E.2: Model estimates of poverty in t, conditional on poverty status in $t-1$: African respondents only

Covariate at $t-1$	Poor at $t-1$		Non-poor at $t-1$	
	Marginal effect	|t-ratio|	Marginal effect	|t-ratio|
Individual				
Age	-0.002	(11.91)	-0.003	(4.30)
Age squared	0.00005	(9.55)	0.00006	(3.22)
Female	0.037	(8.08)	0.045	(2.82)
Household head				
Age	-0.001	(0.40)	-0.004	(1.30)
Age squared	-0.00001	(1.45)	0.000004	(0.11)
Female	0.023	(4.61)	0.060	(3.52)
Matric and above	-0.115	(14.92)	-0.231	(11.12)
Employed	0.022	(3.58)	-0.047	(2.29)
Household				
Urban	-0.055	(9.70)	-0.032	(1.59)
Farm	-0.018	(1.72)	0.006	(0.17)
Adult 65 and above	0.007	(0.91)	0.006	(0.17)
Children 15 and below	0.034	(22.24)	-0.007	(0.88)
Any workers	-0.053	(9.34)	0.087	(2.56)
Own dwelling	0.049	(7.31)	0.020	(0.96)
Log likelihood	-72 960			
Model chi-squared (d.f. = 52)	8 644 ($p<0.000$)			
No. of clusters	10 753			
No. of observations	33 199			
No. of observations (person-waves)	71 643			

Source: Own calculations from the first four waves of NIDS. Reference categories for binary covariates: male, male household head, household head does not have matric, household head is not employed, Western Cape province, rural area, no adults over 65 in the household, no children under 15 in the household, no workers in the household, household members do not own the dwelling. The base wave is wave 1.

3.F Probit model of poverty dynamics

Table 3.F.1: Probit model of poverty in t, conditional on poverty status in $t-1$

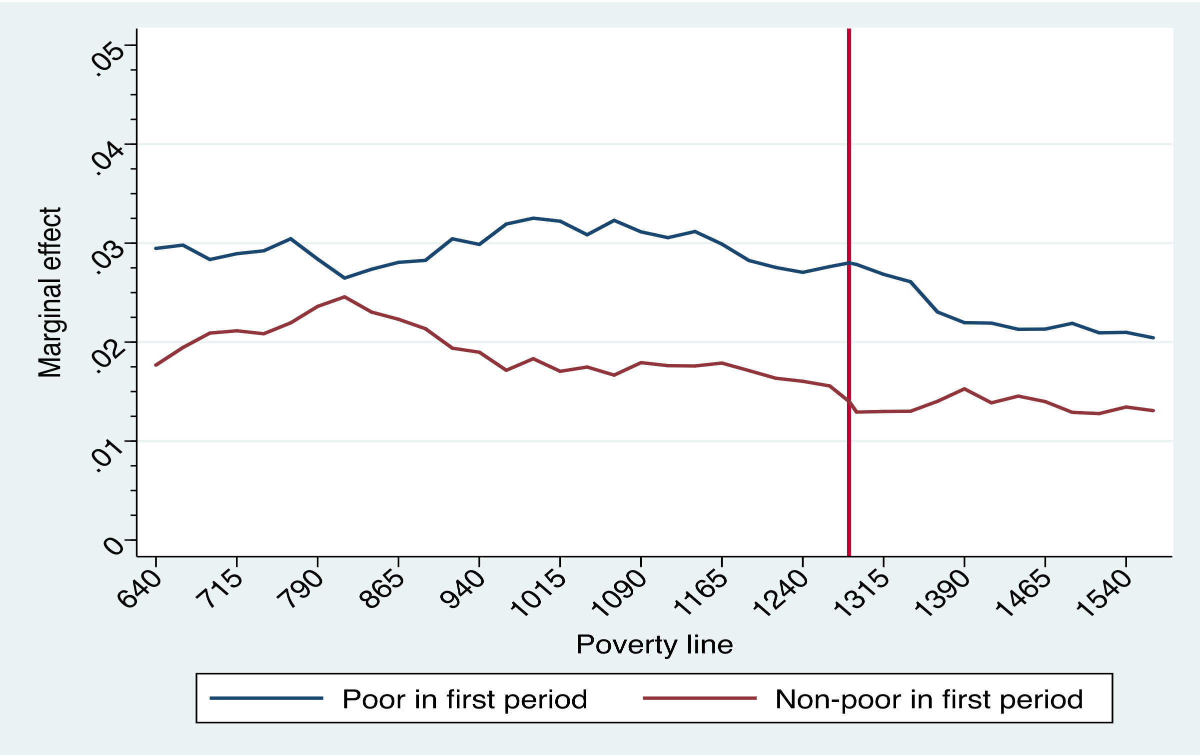
Covariate at $t-1$	Poor at $t-1$		Non-poor at $t-1$	
	Marginal effect	|t-ratio|	Marginal effect	|t-ratio|
Individual				
Age	-0.003	(12.540)	-0.001	(3.280)
Age squared	0.00003	(9.130)	0.00001	(1.520)
Female	0.028	(9.370)	0.014	(2.790)
African	0.235	(6.080)	0.220	(10.830)
Coloured	0.227	(5.830)	0.167	(7.770)
Household head				
Age	0.000	(0.060)	-0.006	(3.090)
Age squared	-0.00001	(0.830)	0.00004	(1.810)
Female	0.018	(2.880)	0.032	(3.390)
Matric and above	-0.125	(13.420)	-0.157	(14.910)
Employed	0.001	(0.100)	-0.021	(1.700)
Household				
Urban	-0.052	(6.590)	-0.038	(2.840)
Farm	-0.001	(0.070)	-0.003	(0.140)
Adult 65 and above	0.006	(0.570)	0.051	(2.670)
Children 15 and below	0.032	(15.380)	-0.031	(2.010)
Any workers	-0.053	(7.000)	0.002	(0.300)
Own dwelling	0.027	(3.530)	0.000	(0.010)
Log likelihood	-28 877			
Model chi-squared (d.f. = 54)	6 142 (p<0.00)			

Source: Own calculations from the first four waves of NIDS. Reference categories for binary covariates: male, white, male household head, household head does not have matric, household head is not employed, Western Cape province, rural area, no adults over 65 in the household, no children under 15 in the household, no workers in the household, household members do not own the dwelling. The base wave is wave 1.

3.G Difference in marginal effects for probit model

3.G.1 Female marginal effects

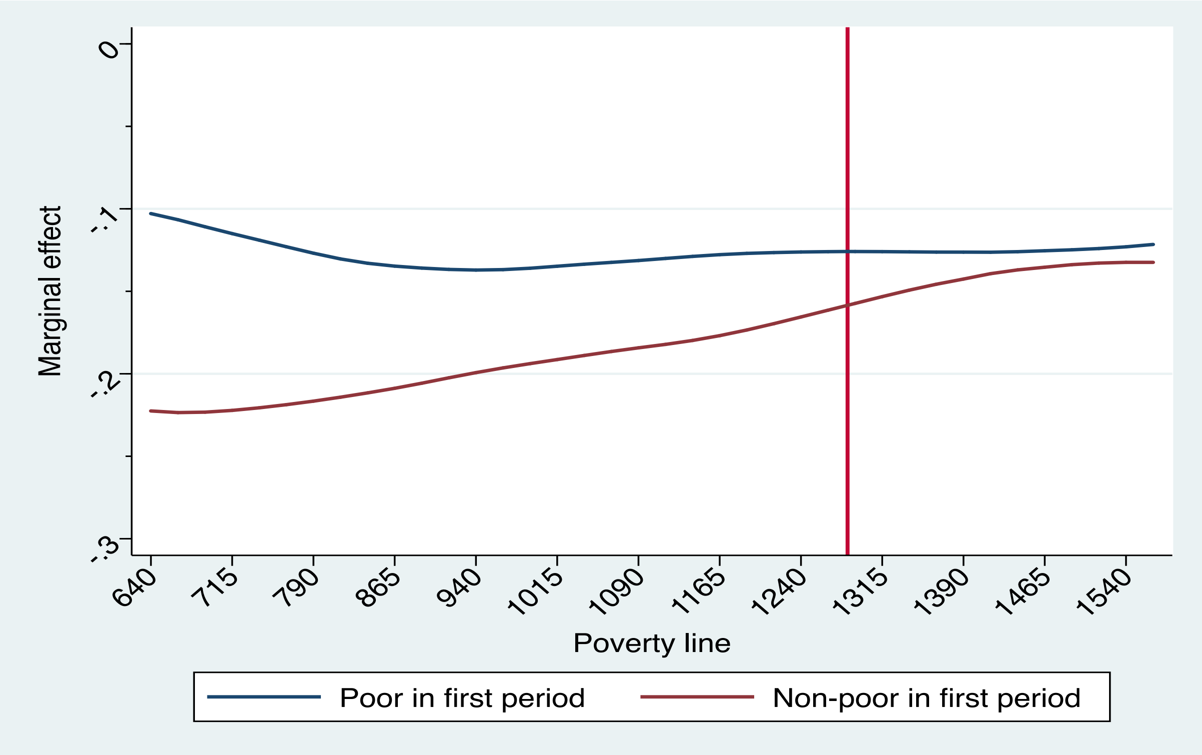
Figure 3.G.1: Difference in female marginal effect for different poverty lines: Probit model only



Source: Own calculations from the first four waves of NIDS.

3.G.2 Household head with at least a matric marginal effects

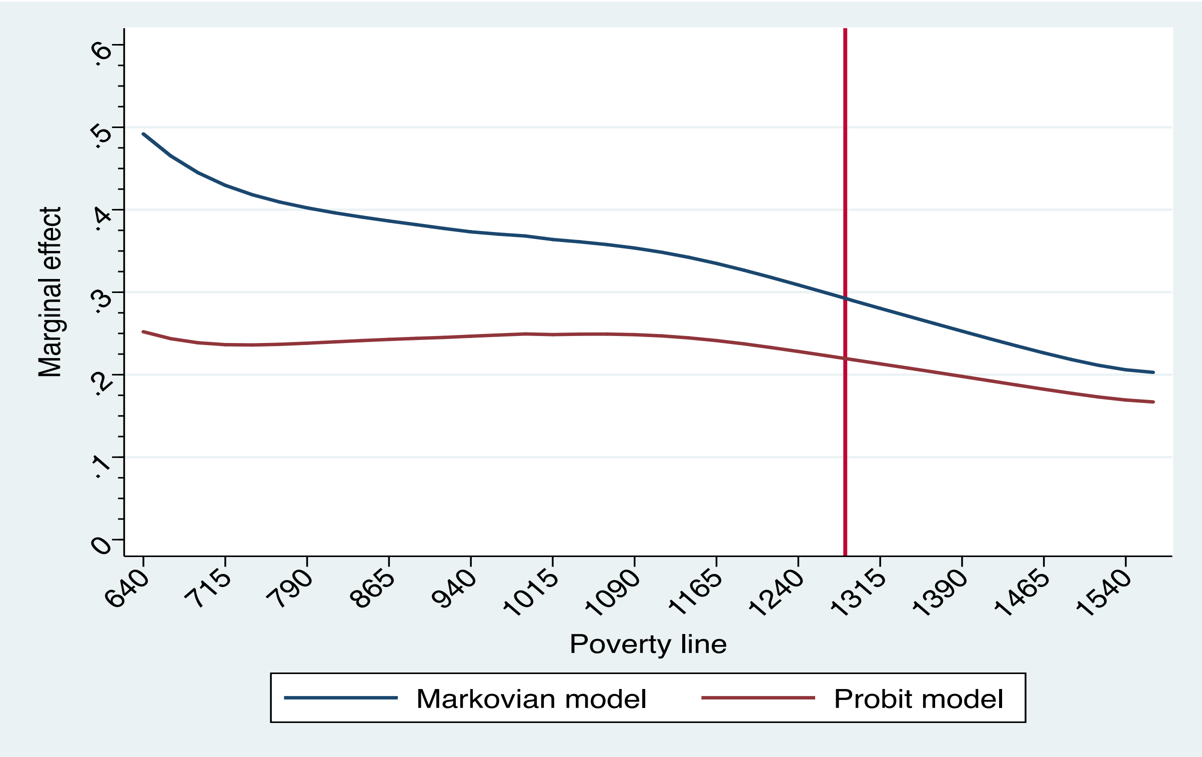
Figure 3.G.2: Difference in household head with matric or above marginal effect for different poverty lines: Probit model only



Source: Own calculations from the first four waves of NIDS.

3.G.3 African marginal effect: Markovian model compared to probit model

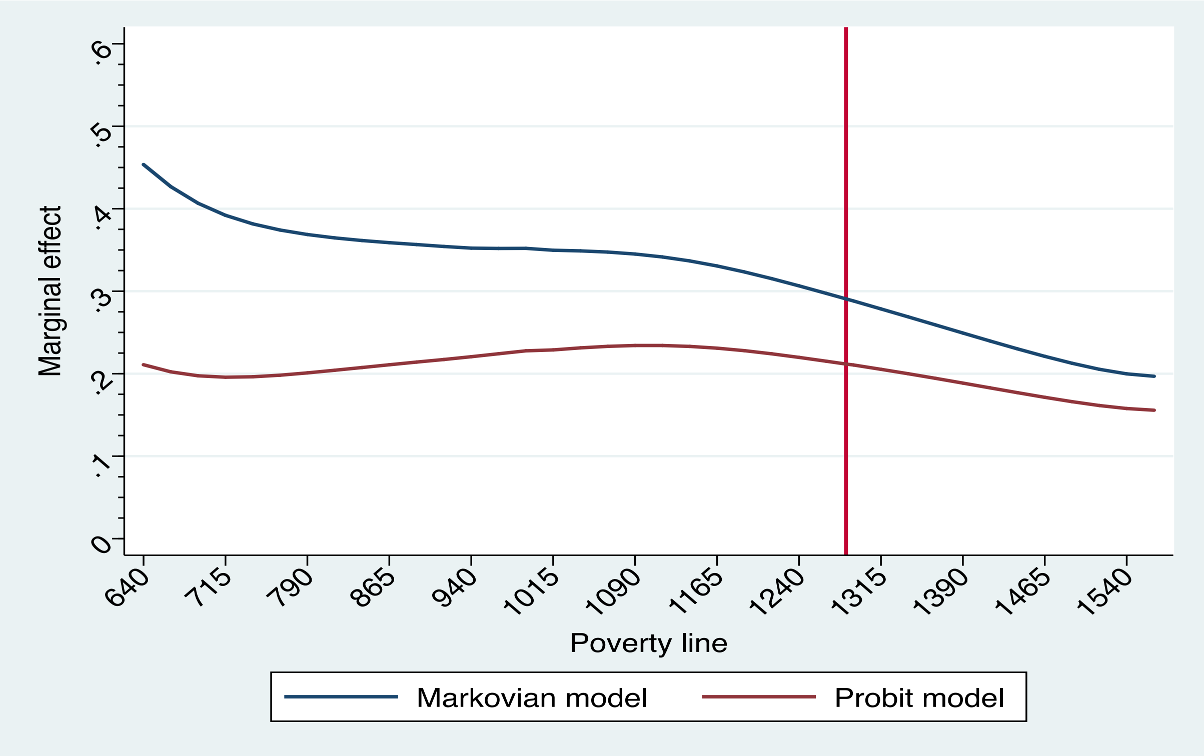
Figure 3.G.3: Difference in African marginal effect for different poverty lines: Markovian model compared to probit model



Source: Own calculations from the first four waves of NIDS.

3.G.4 Coloured marginal effect: Markovian model compared to probit model

Figure 3.G.4: Difference in Coloured marginal effect for different poverty lines: Markovian model compared to probit model



Source: Own calculations from the first four waves of NIDS.

4 Patterns of persistence: Intergenerational mobility and education in South Africa

4.1 Introduction

South Africa has long been highlighted as a country with some of the highest cross-sectional inequality in the world. Studies of why the level of disparity in economic outcomes has remained consistently high have touched on many areas, but it is only the recent emergence of high quality longitudinal data that has allowed researchers to begin to unpack the role of intergenerational persistence of income and earnings in shaping longer run trends. The dynamic relationship between the earnings of parents and the earnings of their offspring shapes the unfolding series of snapshot estimates of inequality that have been calculated for the country. Understanding the mechanisms behind these dynamics is therefore an important part of understanding why inequality in South Africa has remained so high.

The degree of persistence of intergenerational earnings is often closely linked to the question of the equality of opportunity present in society. Recently, Corak (2013) has led the cross-country research into this relationship and has produced what has become popularly known as the 'Great Gatsby curve'. This curve shows a strikingly positive relationship between the persistence of earnings from parents to children, and the level of inequality in a country. The implication is that the closer the correlation between parental and child earnings, the higher the level of inequality in society. The corollary is that equality of opportunity is lower in societies with high persistence between the earnings of parents and those of their children compared to societies with relatively lower levels of persistence. Piraino (2015) has undertaken the most comprehensive work in this area using South African data, and has calculated the intergenerational earnings elasticity (IGE) and an inequality of opportunity index for the country. He finds that the level of persistence between the earnings of fathers and sons is very high and is comparable to other developing countries with high levels of income inequality. He locates South Africa along the 'Great Gatsby curve' as a country with both a high level of intergenerational persistence and a high level of economic inequality.[1]

This chapter makes a number of contributions to the literature on intergenerational mobility,

[1] A statistical link may be drawn between level of inequality of opportunity and the degree of intergenerational persistence in wages in a society. Piraino (2015) shows that when the IGE is estimated using TSTSLS methods, the two concepts are linked. This is because the larger the inequality in parental earnings across different 'types' (defined on observable characteristics), and the higher the intergenerational elasticity of earnings, the higher inequality of opportunity in society will be.

with a particular focus on South Africa. First, it examines how to estimate and analyse intergenerational earnings mobility in a society that has experienced consistently high unemployment over a number of decades. Existing estimates of the intergenerational earnings elasticity have implicitly assumed that selection into employment plays no role in driving the relationship of earnings between parents and their children. In this chapter we estimate the IGE using a double correction which accounts for the high unemployment rates in both generations, and has substantial impacts on the estimated IGE. Second, this chapter uses quantile regressions to investigate non-linearities of the IGE, and presents the first evidence of the shape of the relationship between the earnings of parents and the earnings of their children over the full earnings distribution. This has important implications for how we think about poverty and inequality traps in South Africa. Third, the chapter presents the first estimates of the IGE of mothers relative to sons, in order to provide a fuller picture of the intergenerational transmission of earnings. Fourth, the estimation in this chapter takes the issue of non-random attrition very seriously, and a set of panel weights for use over the first four waves of NIDS is created so that we are able to construct a comparable sample of sons with earnings over the longest possible period in the data. Finally, the chapter analyses the role of education in driving the intergenerational transmission of earnings. This is done by uncovering the impact of education on the intergenerational elasticity of earnings over the full earnings distribution, and then by decomposing the IGE into relative contributions of education versus skill. The first, second and fifth contributions are extensions to the literature in their own right, while the third and fourth contributions can be thought of as direct extensions of Piraino (2015).[2]

The structure of this chapter is as follows. Section 4.2 presents three stylised facts about the South African economy and labour market which have motivated this chapter. Section 4.3 discusses the relationship between intergenerational mobility and inequality, and outlines the theoretical framework that will be used to measure and decompose the intergenerational earnings elasticity. Section 4.4 describes the data and estimation procedures used in our study, and presents some descriptive statistics. In section 4.5 we report the results from a number of different estimations of the intergenerational earnings elasticity, and this is followed by an

[2]Piraino (2015) also presents an analysis of inequality of opportunity in South Africa, as well as a racial decomposition of the IGE, neither of which form part of this chapter.

analysis of the role of education in determining and shaping this elasticity. The final section provides some concluding remarks.

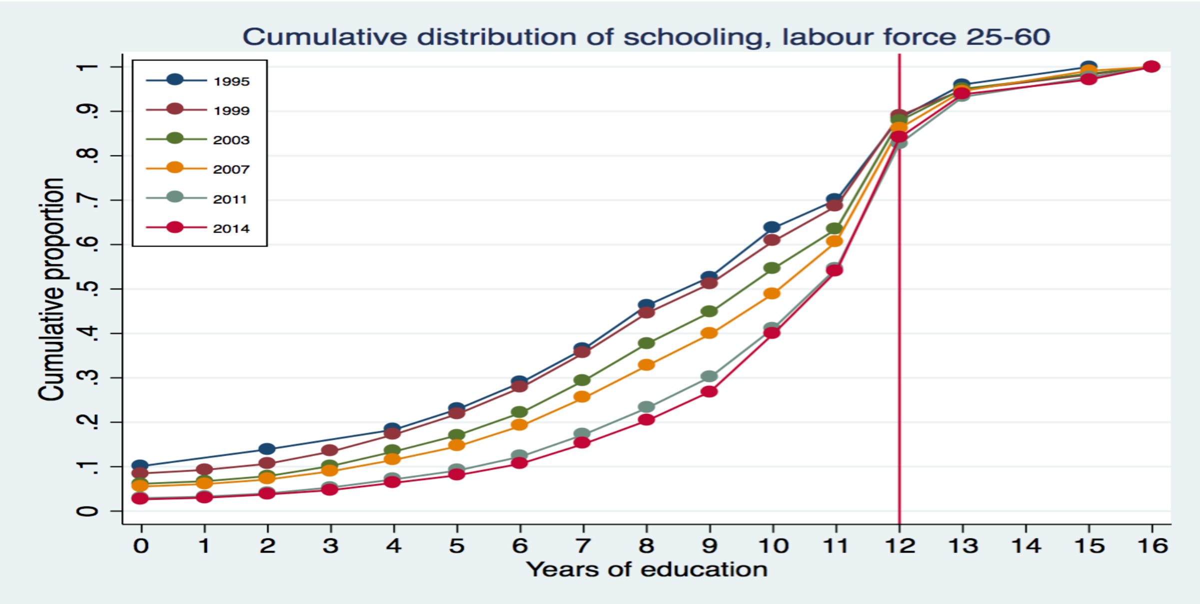
4.2 Motivation for the study

This chapter is largely motivated by three stylised facts that have emerged from the post-apartheid South African literature on education and economic inequality. First, there has been an increase in the general level of educational attainment, along with a reduction in the inequality of education levels. Second, the last two decades have seen an increase in the returns to matric and post-matric education relative to other education categories. Third, the levels of cross-sectional economic inequality and unemployment have been very high and persistent in South Africa's democratic era.

4.2.1 Stylised fact 1: Increased educational attainment

There has been a sharp decrease in the inequality of educational attainment in the country, and this has come about because of a general increase in the number of years of schooling completed by South Africans. The coefficient of variation of education for working-age South Africans fell from 0.5 in 1994 to just over 0.3 in 2011 (Lam et al., 2015). The increase in educational attainment for working-age South Africans is confirmed in Figure 4.1 which is reproduced from Lam et al. (2015), with the addition of data from 2014. Improvements in the average level of education are evident in the cumulative distribution functions (CDFs) from 1995 to 2014, with the increase being driven by higher proportions of the labour force completing primary school. In 1995 more than half the labour force had dropped out of school by grade 9. By 2014 this proportion was below 30%. Though the increase in educational attainment is impressive, a figure of CDFs remains agnostic as to the quality of that education. The question of quality is analysed in van der Berg et al. (2011), who show that low quality education in South Africa is a poverty trap, the ill effects of which are borne disproportionately by pupils attending historically black schools. Branson et al. (2014) show that school dropout in South Africa is largely driven by falling behind (defined as being more than two years older than expected for the current grade), even after controlling for socio-economic factors. Falling behind is itself determined largely by school quality, with historically black schools lagging particularly far behind in this regard. This reiterates the findings in Ardington et al. (2011) who show stark racial differences in progress through school using data from the Cape Area Panel Study (CAPS).

Figure 4.1: The increasingly educated South African labour force



Source: 1995 to 2011 based on Figure 1 in Lam et al. (2015) using the Post Apartheid Labour Market Series (PALMS) dataset. 2014 calculated using the Labour Market Dynamics in South Africa (LMDSA) 2014 dataset.

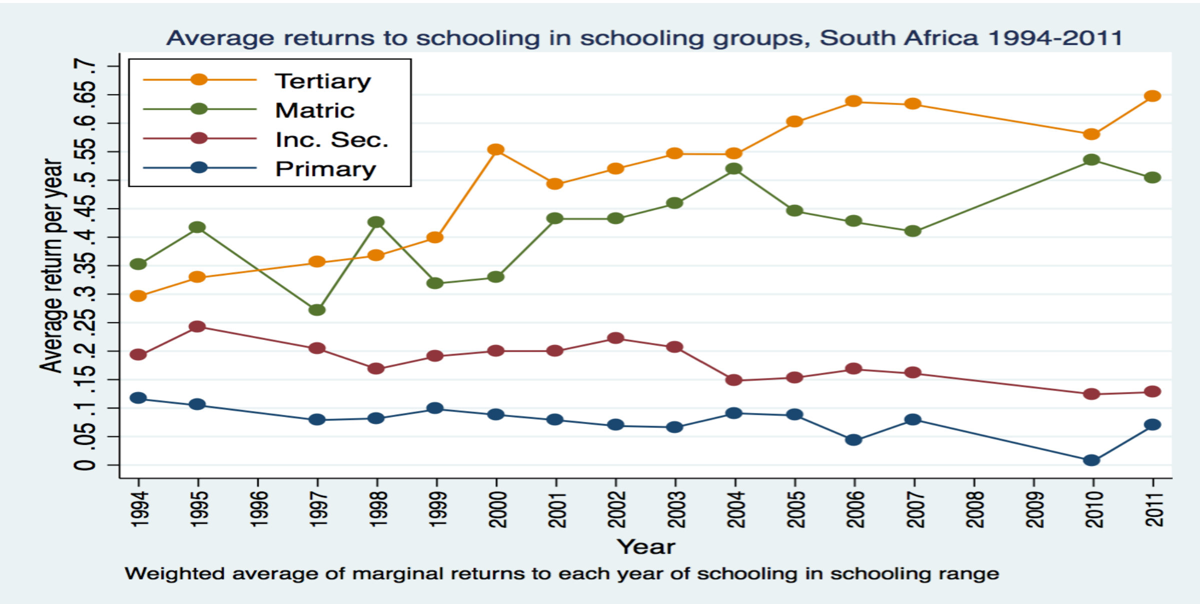
4.2.2 Stylised fact 2: Changing patterns in the returns to schooling

Although the average level of education of the South African labour force has increased, this may not have been matched by a proportional increase in earnings. Keswell and Poswell (2004) and Branson and Leibbrandt (2013) among others have found that the country displays a strongly convex returns to education function, even once experience and educational quality are controlled for. Figure 4.2 below is adapted from Lam et al. (2015)[3] and plots the average returns to schooling for four schooling groups from 1994 to 2011. The returns are calculated as the weighted average of the marginal returns to an additional year of schooling for each

[3]Lam et al. (2015) plot only three categories – primary, incomplete secondary, and matric and above.

year within the ranges of primary, incomplete secondary, matric and tertiary. The figure shows that there has been an increase in the returns to matric and post-matric education relative to the incomplete secondary and primary schooling categories. The increase in the returns to matric and above occur at the same time as a relative decrease in the returns to both of the other categories. The full benefits from a more educated labour force are therefore not translated into a proportional increase in earnings unless a worker has completed high school and continues into postsecondary education. This resonates with concerns about the persistent nature of inequality, as Corak (2013) notes that relatively higher returns to tertiary education often go hand-in-hand with high and sticky cross-sectional inequality.

Figure 4.2: Returns to schooling by schooling category



Source: Based on Figure 1 in Lam et al. (2015) using the Post Apartheid Labour Market Series (PALMS) dataset.

4.2.3 Stylised fact 3: Stubbornly high economic inequality

Although there is some debate as to the precise level of economic inequality in South Africa, there is no doubt that it has been consistently high in the post-apartheid period. The Gini coefficient for labour market earnings in South Africa has averaged around 0.55 (Finn, 2015), while the Gini coefficient for total household income per capita has been at 0.66 or above since 1993 (Leibbrandt et al., 2010; Yu, 2010). Although the level of inequality has remained high in the post-apartheid period, one important change is that the relative weight of inequality between races has decreased, while the importance of inequality within racial groups has risen steadily (Leibbrandt et al. (2012), Finn (2015)). Part of the blame for the stickiness of inequality in South Africa comes from the dynamics of how educational attainment and labour market earnings of parents feeds through to the educational attainment and earnings of children. Uncovering part of this dynamic relationship is an important part of understanding inequality in contemporary South Africa, and is the main contribution of this study.

4.3 Theoretical background

There has recently been something of a shift in the focus of the inequality literature, with studies of inequality of opportunity becoming more prevalent relative to studies of the inequality of outcomes. A key feature of these works is the attempt to distinguish between inequality that arises because of inherited circumstances and inequality that arises due to the application of effort. The former, which is often subsumed in the idea of inequality of opportunity, is usually seen as less ethically justifiable than the latter. If variables that are beyond a person's control, such as parental education, race or sex, do not have any bearing on their realized economic outcome, then one may say that there is equality of opportunity because differences in economic outcomes are driven by the effort expended by each individual, and by luck. However, as noted by Atkinson (2015) the distinction between inequality of opportunity and inequality of outcomes is not a clear one in either a single generational or intergenerational sense. The reason for this is the fact that 'today's ex-post outcomes shape tomorrow's ex-ante playing field' (p. 11). There is no reason to think that equal opportunities will lead to equal outcomes in a dynamic sense. Even if it were possible for an entire generation to start off with identical opportunities, the unequal ex-post distribution of economic outcomes would mean that the next generation would face ex-ante inequality.[4] If the starting point for each generation is highly unequal, and the transmission of economic outcomes from parents to children is largely deterministic, then this has clear implications for the persistence of inequality in society. Therefore South Africa's, low level of *intergenerational* mobility has dynamic consequences for the pro-

[4]This presumes that the state is non-interventionist in equalizing ex-ante opportunities.

duction of *intragenerational* inequality, and understanding this relationship is important from a policy and ethical perspective.

The focus on labour market earnings is warranted because of the important role that wages play in determining the extent of cross-sectional inequality in South Africa (Leibbrandt et al., 2010). A better understanding of wage inequality goes a long way to assist an understanding of household income inequality, and understanding intergenerational earnings mobility goes a long way to explaining why inequality has been so persistent in South Africa.

The reason for focusing on education as a transmission mechanism is because education is widely cited as being the key factor in reducing cross-sectional inequality, but an equalization of education may not lead to lower inequality, as we have witnessed over the last twenty years in South Africa. The impact of equalizing education, therefore, cannot be seen in isolation. It must be understood together with the labour market outcomes associated with education. These include the probability of finding a job and the shape of the returns to education function itself.

Another dynamic to note is that credit constraints may be significant barriers to both the quantity and quality of education a child receives, and this can contribute to a pattern of inequality that is self-reinforcing. Furthermore, the higher the correlation of economic outcomes between parents and children, the longer it takes for a society to reach the equilibrium social status of each generation (Checchi, 1997). The introduction of public policies that lower the explicit and implicit costs of public education, along with those that improve quality in order to ease the transition from one level of education to the next, are therefore crucial factors in increasing the intergenerational mobility of economic outcomes.

A seminal theoretical paper by Becker and Tomes (1979) sets about trying to explain the dynamics of educational attainment from generation to generation.[5] One of the central motivations of this paper is to unify the analysis of cross-sectional inequality (inequality within a generation) and intergenerational inequality. The persistence of income from one generation to the next is determined by a mix of factors including the level of endowments of an individual, the inheritability of various characteristics, the propensity of each generation to invest, and luck.

[5] An even earlier intergenerational transmission model along with its implications for social mobility was proposed by Conlisk (1974), though it has received less attention in the social mobility literature.

The Becker and Tomes (1979) framework has inspired a large body of economic theory on the transmission of economic advantage between generations that is distinct from the sociology literature which preceded it by several decades. Empirical applications inspired by these models soon followed and, as noted in Chusseau et al. (2013), one of the defining features of this literature is the attempt to separate out the roles of 'effort' and 'luck' in determining social mobility, and this is generally done by isolating the influence of different channels that determine educational attainment and labour market returns.

Lefranc and Trannoy (2005) present a simplified version of the Becker and Tomes model. Let us assume that the transmission of income or earnings from parent to child is determined by the individual endowment of human capital, and by the innate ability of the child. The Becker and Tomes model is built on the assumption that the child's utility enters the parent's utility function, and that the child's level of human capital is chosen by the parent as a result of the optimal allocation of permanent income. The relationship between the child's permanent income (denoted by c) and the parent's permanent income (denoted by p) is given by the following equation:

$$Y^c = \phi Y^p + \theta a^c \tag{4.1}$$

In this equation the parameter ϕ represents the extent of the causal relationship between the permanent income of the parent and the permanent income of the child. As noted by Lefranc and Trannoy (2005), the source of this correlation maps to the positive relationship between the father's earnings and the investment in the child's human capital. The constraint on this investment is the amount of financial resources available to the family. This is something that may be particularly important in South Africa, as credit constraints have been shown to be a barrier to postsecondary enrollment (Lam et al., 2013). In addition, the constraint may bind before postsecondary enrollment by limiting parents' ability to send their children to a better school that may require a higher level of expenditure.

The second term on the right hand side captures the determinants of the child's permanent income that are related to factors that 'money can't buy'. These include things like IQ, social networks, or preferences (Lefranc and Trannoy, 2005). Becker and Tomes (1979) differentiate this effect from the previous effect by noting that its influence on the intergenerational transmission of income comes from earnings determinants that are independent of parental investment decisions.

Separating out the two different types of transmission mechanisms that arise from the Becker and Tomes model would yield interesting policy implications. If the dominant mechanism determining intergenerational earnings transmission is parental investment in education, then overcoming credit constraints would lead to a smaller correlation between the earnings of successive generations, and therefore more economic mobility. If, however, the dominant mechanism is individual ability, then increasing social mobility by weakening the relationship between the earnings of parents and their children may be more difficult.

Much of the research that is motivated by this theoretical model does not make a distinction between the two mechanisms explaining intergenerational earnings. In general, a simple regression of son's permanent income[6] (or earnings) on father's permanent income (or earnings) is the preferred approach, given the data's inability to convincingly isolate the 'ability' mechanism. Combining both mechanisms into a single coefficient will lead to an upward bias in the estimate of the elasticity of intergenerational earnings (Lefranc and Trannoy, 2005). In this chapter we estimate the intergenerational elasticity of earnings using a reduced form version of this model, in line with most of the international literature. In doing so we attempt to overcome the bias that may arise from co-resident selection, and the bias that may arise from selection into a job in a society with a very high unemployment rate.

The canonical estimation of intergenerational mobility comes from a simple regression of the logarithm of child's (usually son's) permanent income on the logarithm of parent's (usually father's) permanent income.

$$Y_i^c = \alpha + \beta Y_i^p + \varepsilon_i \tag{4.2}$$

β is generally referred to as the intergenerational elasticity of earnings (IGE), when labour market returns are the focus, and is the most commonly used measure of the persistence of

[6]Most of the studies in the international literature focus on the correlation of earnings between fathers and sons because of the added complication of accounting for female labour market participation decisions.

earnings between generations. As β is a measure of persistence, $(1 - \beta)$ may be thought of as a measure of intergenerational mobility. As β approaches zero, society approaches a situation of perfect intergenerational mobility in which the earnings of the parent do not determine the earnings of the child. Conversely, as β approaches 1, the earnings of the parent increasingly determine the earnings of the child, and intergenerational mobility goes to zero. Though the interpretation of the intergenerational elasticity in this model cannot be interpreted in a purely structural sense, it is nonetheless a widely used and useful descriptive measure of how persistent earnings are between generations.

Another descriptive statistic that has been used widely in the literature is the intergenerational earnings correlation, ρ. As shown in Jäntti and Jenkins (2013) the relationship between the β measure of intergenerational earnings elasticity and the Pearson product moment correlation is given by the following:

$$\rho = \beta \frac{\sigma_{y_p}}{\sigma_{y_c}}$$

where σ_{y_p} and σ_{y_c} are the standard deviations of log earnings in the child's generation and the parent's generation respectively. This measure also highlights the link between intergenerational mobility and inequality, as the numerator and denominator on the right hand side are the log variance inequality indices for the parent's and child's generations respectively.

The intergenerational elasticity measure has been preferred to the intergenerational correlation in much of the literature for a number of reasons. First, as noted in Lefranc and Trannoy (2005), the elasticity may be measured independent of calculating the inequalities in each generation. Second, intergenerational elasticity is perhaps a more intuitively appealing concept to economists than the intergenerational correlation. Consider a policy shift that reduces the deviation from the mean of all income in the child's generation by the same factor. The effect of this policy should see a decrease in the persistence of intergenerational income (that is, an increase in intergenerational mobility). Indeed, the intergenerational elasticity would decrease under this policy, but the intergenerational correlation would not. The intergenerational correlation would remain unchanged, and the increased mobility would not be reflected. Third, the intergenerational elasticity is not biased if there is measurement error in the variable reflecting child's earnings (the dependent variable in the regression), unlike the correlation (Black and Devereux, 2011). Finally, as Jäntti and Jenkins (2013) point out, researchers may want to compare their estimates of intergenerational mobility to those of other studies, and the popularity of the measure ensures its continued use independent of any theoretical concerns.

Many studies have calculated the intergenerational elasticity of earnings in the last five to ten years. Reviews and international comparisons can be found, among others, in Blanden (2009), Brunori et al. (2013) and Corak (2013), which all provide tables of the intergenerational elasticities for a number of countries. The international evidence lends support to the 'Great Gatsby Curve', which suggests that countries with higher levels of inequality have lower levels of intergenerational mobility. Countries with low levels of cross-sectional inequality - in particular Scandinavian countries - have a higher degree of intergenerational mobility (a lower intergenerational elasticity) than those with a higher degree of inequality such as the United States, the United Kingdom, and Italy (Corak, 2013). The Scandinavian countries have intergenerational elasticities that are below 0.2, while for countries with higher levels of inequality the elasticity is around 0.5.

Intensive data requirements have precluded the calculation of intergenerational elasticities for developing countries until recently. Piraino (2015) notes that these developing countries tend to have less intergenerational mobility than their OECD counterparts, and calculates an intergenerational elasticity that is between 0.57 and 0.67 for South Africa, depending on the variables used in the imputation of father's earnings. In other examples, Hnatovska et al. (2013) calculate an elasticity of around 0.5 for India, while Ferreira and Veloso (2006) find an elasticity of about 0.58 in Brazil. Grawe (2004) calculates an elasticity of 0.54 in Malaysia and 0.67 for Peru, while Bevis and Barrett (2015) calculate separate elasticities for sons and daughters, but find an average of about 0.5 in the rural Philippines. Recent data from urban China put the elasticity at around 0.6 (Gong et al., 2012), though the authors find that intergenerational persistence is far stronger for sons than it is for daughters. Asadullah (2011) calculates an intergenerational wealth elasticity of 0.538 for rural Bangladesh. Recent estimates for Ethiopia (Haile, 2016) and Vietnam (Doan and Nguyen, 2016) calculate intergenerational earnings elasticities of 0.357 and 0.48 respectively.[7]

The research on intergenerational income mobility in South Africa is relatively sparse, with the first example being Hertz (2001) who uses data on co-resident fathers and sons in the KwaZulu-Natal Income Dynamics Study (KIDS) to calculate a range of intergenerational elasticities. A problem facing any analysis of this kind is the fact that the co-residency requirement may introduce selection bias into the estimation – the wages and characteristics of sons who co-reside with their fathers may be different to the wages and characteristics of those who do not.

A number of studies address this concern by making use of a two sample two stage least square (TSTSLS) estimation in which the earnings of the fathers are imputed using a nationally representative dataset from a previous time period. Piraino (2015) adopts this method and locates South Africa's position on the 'Great Gatsby Curve', adding further evidence to the pattern of high-inequality societies having low intergenerational mobility. He also links the literature on intergenerational mobility to that focusing on the inequality of opportunity, and finds that South Africa's inequality of opportunity index is high by international standards. We use Piraino's approach as a benchmark in our calculation of the intergenerational income elasticity, and build on this to highlight the role of education in shaping the earnings dynamics from generation to generation.

4.4 Data and estimation procedure

Calculating the intergenerational elasticity of earnings and extracting the contribution of education to this elasticity requires data that are not often present in a single dataset. The ideal dataset would be a long panel that allows the researcher to calculate permanent income for both parents and children, whether they co-reside or not.

Given that this kind of comprehensive dataset is not yet available in South Africa, in this

[7]Clean comparisons of the estimates of intergenerational earnings mobility between different countries are difficult, as a variety of estimation methods are used, along with different measures of income. Estimates in OECD countries tend to use either long panels (see, *inter alia*, Jäntti et al. (2006) for Denmark and Schnitzlein (2016) for Germany) or large administrative or tax databases (see, *inter alia*, Nybom and Stuhler (2016) for Sweden and Nilsen et al. (2012)). Estimates using data from developing countries tend to be derived using an instrumental variables approach (India, Philippines and China) or a two-stage approach (Brazil, Malaysia, Peru, South Africa and Vietnam), and may use earnings, predicted earnings, income or wealth as measures of economic welfare.

study we make use of two different datasets that allow us to calculate the earnings of two generations. Earnings for the second (younger) generation are calculated using the first four waves of the National Income Dynamics Study (NIDS) which were collected in 2008, 2010/2011, 2012 and 2014/2015, respectively. NIDS contains comprehensive information about the labour market activities and earnings of adults in the sample. Monthly earnings are calculated by combining reported income from all jobs, self-employment activities, profit shares, and bonuses.

One option available to researchers who want to calculate the earnings of the parental generation is to focus on families in which children co-reside with parents. Indeed, this is the approach adopted by Hertz (2001) using data from KwaZulu-Natal. There are at least two significant problems with this approach. First, the subsample of co-resident parents and children may be relatively small. Second, selection bias may be introduced by restricting the analysis to those children who earn wages and still live with their parents. Co-resident children may have observed and unobserved characteristics that are systematically different from those who do not live with their parents, and this will bias our estimates of the intergenerational earnings elasticity.

The adult questionnaire in NIDS asks respondents a series of questions about their parents who are either non-resident or deceased.[8] These include the age, education, and occupation of the parent. Thus, even if a parent is not interviewed directly, we are able to impute the earnings of the parent for a given set of characteristics. Following Piraino (2015) we use nationally representative data from 1993, the Project for Statistics on Living Standards and Development (PSLSD) to generate an earnings variable for the parental generation in NIDS. This Two Sample Two Stage Least Squares (TSTSLS) approach is explained in detail below, but in summary the following takes place. First an earnings regression is run on the PSLSD 1993 data in order to capture the determinants of wages in the parental generation. The dependent variable is the log of wages, and the independent variables are education categories, race, occupational categories and province of residence. These independent variables are chosen because they are the same as those reported by children about their parents in the NIDS dataset. Earnings for parents in NIDS are then imputed by using the estimated coefficients from the wage regressions

[8]Respondents are asked about the highest level of education completed by their mother and father, and about the current or last job in which each parent was employed. Co-resident parents are interviewed directly.

in the 1993 data, along with parental characteristics from the NIDS data. The approach was introduced by Klevmarken (1982) and is sometimes thought of as a ‘cold deck’ linear regression imputation because an auxiliary sample is used to impute the missing variable of interest in the main sample.

Drawing on the exposition in Cervini-Plá (2015) and Lefranc and Trannoy (2005), we estimate the TSTSLS variant of the intergenerational earnings elasticity in the following way.[9] In NIDS - what we call our main sample - we have information about Y^c, but not about Y^p. NIDS also contains sociodemographic information about parents contained in the vector of characteristics Z. The auxiliary sample, the PSLSD 1993, contains a wage variable for parental earnings, Y^p, as well as the same vector of characteristics Z.

Let us begin by denoting the log of parental earnings at time t, Y^p_{it} as:

$$Y^p_{it} = Y^p_i + u^p_{it} \tag{4.3}$$

where the error term captures transitory shocks as well as measurement error in parental earnings. We assume that the log of earnings in the child’s generation is related to the log of permanent earnings in the same way, and that the errors from the parental and child generations are not correlated. For the vector of characteristics Z^p_i (in our case education, occupation, race, and province of residence), we assume that current parental income can be written as:

$$Y^p_{it} = Z^p_i \gamma + \upsilon^p_i + u^p_{it} \tag{4.4}$$

in which the time invariant error is uncorrelated with the set of characteristics. Our first problem is that Y^p_{it} is not available in our main sample I, in this case NIDS. However, in the PSLSD 1993, sample J, we have the same nationally representative population as in NIDS and we are able to extract an estimate of γ, $\hat{\gamma}$ which is obtained through estimating parental earnings in the auxiliary sample J:

$$Y^p_{jt} = Z^p_j \gamma + \upsilon^p_j + u^p_{jt} \tag{4.5}$$

[9]The TSTSLS method for calculating the intergenerational income elasticity first appeared in Björklund and Jäntti (1997).

in which $j \in J$. This is then used to form a prediction of parental earnings in the main sample, and in turn estimate β in the following way:

$$Y_{it}^{c} = \alpha + \beta(Z_{i}^{p}\hat{\gamma}) + \eta_{it} \tag{4.6}$$

Björklund and Jäntti (1997) note that if the characteristics in Z are also determinants of the child's income, then the intergenerational earnings elasticity will be biased upwards. That is, if the parental level of education and occupational category both have a positive impact on the child's earnings, then the elasticity may be biased upward. In this light, many calculations of the intergenerational elasticity using the TSTSLS method can be thought of as an upper bound for the measure of income persistence between generations (Piraino, 2015). At this point it is useful to echo Blanden (2015) who stresses that the intention is not necessarily to extract the causal effect of parental income on child income. Rather, the intention is to generate a measure of persistence of earnings across generations in a similar vein to how the Gini coefficient measures inequality in a cross-section.[10]

In this study we estimate the intergenerational earnings elasticity using equation 4.6. Bootstrapped standard errors from 500 repeated processes are reported in which parents and children are resampled separately, and both stages of the model are estimated in each repeat of the bootstrap. We also adjust for the fact that we observe both parents and children at different stages of their age-earning profiles using the method outlined in Bratberg et al. (2007). This is done separately for parents and children by regressing earnings on age and age squared, and then using the sum of the constant term and the residual from that regression as the measure of earnings. The age range for the younger generation is from 20 to 44 years old,[11] while for the parental generation in the PSLSD 1993 dataset, we focus on adult earners between 30 and 59 years of age.

In general the focus will be on the relationship between the earnings of fathers and sons, and mothers and sons. This is in line with most of the international literature which avoids

[10]Blanden (2015) draws the inequality-mobility connection succinctly in saying, 'The Gini coefficient provides a summary measure of cross-sectional inequality, but it does not provide any information about its source. The intergenerational elasticity measure performs a similar function for intergenerational inequality.'

[11]The age interval refers to the age of the respondent in the first wave of NIDS.

parent-daughter estimations due to the added complication of adjusting the elasticity to account for female labour market participation decisions.[12] In this chapter we acknowledge the difficulty of correcting for the bias that may arise from this process, and report both father-son and mother-son elasticities for the most part. We also evaluate how sensitive our measures of intergenerational mobility are in a high-unemployment labour market, and report the selection-adjusted measures for both fathers and mothers.[13]

We restrict our analysis to those sample members who appear in all four waves of NIDS as this allows us to get a measure that is as close to permanent income as possible, given the data constraints.[14] Averaging earnings across four waves for the second generation will get us closer to this than using single points in a cross-section. We are unable to estimate a similar averaged measure for the parental generation, and instead use the single imputed earnings point, as described above.[15] Another reason for choosing to focus on the balanced panel members is the fact that we are able to correct for selective attrition by using panel weights. Attrition rates between each of the waves of NIDS varied widely by racial group and by socio-economic characteristics (de Villiers et al., 2013). White respondents were more likely to drop out between waves than any of the other racial groups, as were those who were relatively wealthier. We construct attrition-corrected longitudinal weights in the same way as Finn and Leibbrandt (2013). This involves modelling attrition by running a series of unfolding probit models from wave 1 to wave 2, from wave 2 to wave 3, and from wave 3 to wave 4. The wave 2 longitudinal weight is constructed by multiplying the wave 1 post-stratified weight by the inverse of the conditional probability of re-interview in wave 2. The same process is applied

---

[12] The parental sample is drawn from data from 1993, just before a surge in female labour force participation rates in South Africa. Female labour force participation rates have changed substantially in the post-1993 period (Casale and Posel, 2002). Much of the international literature in fact excludes mother's earnings altogether. We choose to contrast the intergenerational elasticity relative to both father's and mother's earnings as the high father absenteeism rate in South Africa may mean that recalled information about mothers in the sample is more reliable. In practice our sample sizes for mothers are relatively smaller, and the qualitative findings are similar whether fathers or mothers are used.

[13] These are corrected using Heckman's two-step approach, and are reported in section 4.5.

[14] Piraino (2015) pools the data across three waves and uses observations that appear once, twice or thrice in the data. Like Piraino (2015) we also use average earnings in cases where respondents report earnings in multiple waves of NIDS. Each respondent's total earnings is divided by the number of waves in which he age eligible to appear in the sample. If, for example, a respondent appears in all four waves and is aged between 20 and 44 in all four waves, but only reports earning in three waves, then his total earnings will be divided by four, rather than three.

[15] Using snapshots of parental earnings when estimating the intergenerational elasticity biases the elasticity downwards because of the presence of measurement error (for example see Solon (1992).

between wave 2 and wave 3, and between wave 3 and wave 4. The final longitudinal weight is applied to all respondents who were successfully interviewed in all four waves of NIDS. Our sample size is always larger than 1 200, and so we are not overly concerned about power issues given our decision to focus on the balanced panel members. All subsequent analysis in this chapter makes use of this weighting structure.

Table 4.1 presents some descriptive statistics for the balanced panel members that form part of our analysis. The sample is restricted to those males between the ages of 20 and 44 who report their earnings and who have non-missing information about their parents.

The mean age in wave 4 of the 1 785 respondents in our analysis sample is 35. About 85.5% of those in the sample are African, and the proportion of White and Coloured respondents is similar. The second panel of the table presents the proportion of respondents and their parents in different education categories. Consistent with the pattern in Figure 4.1, there is a significant increase in the level of education attainment from parents to children in the sample. Over 40% of respondents reported having parents who had no education, while the corresponding figure for respondents themselves was under 3%. The bulk of the shift in education attainment was to matric and postsecondary education. 43% of respondents in the balanced panel reported having attained at least a matric. The corresponding proportions for the fathers and mothers of these respondents are 12.4% and 11.7% respectively.

In the next panel of the table we present the proportion of respondents, fathers and mothers in different occupational categories. These are based on the South African Standard Classification of Occupations (SASCO) conventions and are adjusted so as to overlap directly with the occupational categories present in the PSLSD 1993 data. These categories can be thought of as loose proxies for occupational skill level, and mirror those used by Keswell et al. (2013) in their study of intergenerational occupational mobility in South Africa. The categories are rather broad and in reality each category probably covers a wide range of skill levels itself, but they are reported here as they form part of the imputation for parental earnings in the first stage of the TSTSLS estimation. The occupational distributions for fathers and sons look relatively similar, though there are a higher proportion of sons in clerk/sales categories, and a lower proportion in operator/semi-skilled jobs than their fathers. Almost 60% of the mothers of our balanced panel respondents were employed in elementary occupations - nearly three times the proportion of sons. Interestingly, 13.2% of mothers were employed in the highest skill category (professional/technical/manager), and the corresponding percentage for sons was lower at 8%.

Direct comparisons of unemployment rates over time in South Africa are not possible, as noted in Kerr and Wittenberg (2016), because StatsSA changed the definition of what is considered work, as well as the criteria for being considered to be searching for employment. Bearing these reservations in mind, the table presents unemployment rates for fathers and mothers in 1993, and for sons in 2008 and 2015. These are shown in order to get a sense of the magnitude of the unemployment problem, rather than to indicate any trends in unemployment. There are two unemployment rates that are generally used in South Africa. One requires active job search in the last 14 days (narrow definition) and one which includes all those who say that they want a job but have not actively searched in the last 14 days (broad definition). Unemployment rates according to the latter definition are shown in the table. It is clear that unemployment was very high in both generations - 26% for males and 33% for females in 1993. The unemployment rate for males during the first wave of NIDS stood at 23%, and by the end of the fourth wave this had risen to 29.6%.

Table 4.1: Summary statistics of the balanced panel

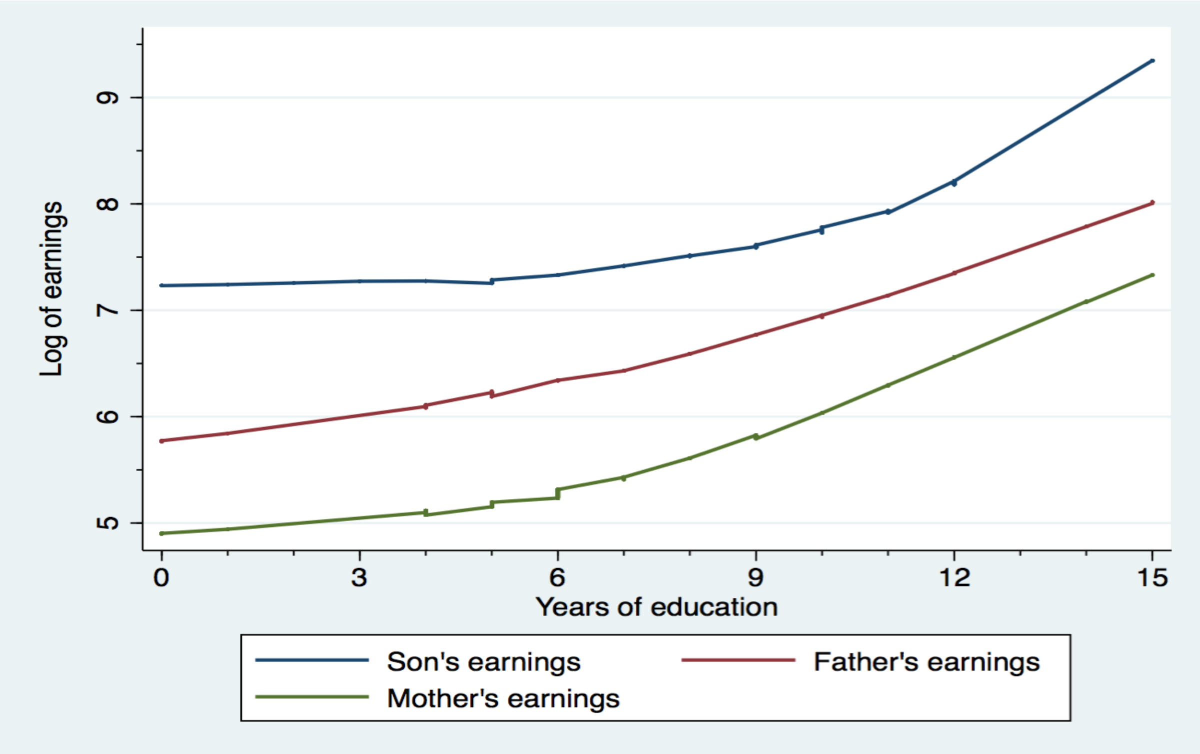
Age (mean in wave 4)	35.41		
Race			
African	85.53		
Coloured	6.49		
Asian/Indian	2.23		
White	5.75		
Education	**Son**	**Father**	**Mother**
None	2.86	46.50	40.39
Primary	13.32	18.54	24.17
Incomplete secondary	40.97	22.62	23.78
Matric	21.90	8.22	6.95
Postsecondary	20.95	4.13	4.71
Occupation	**Son**	**Father**	**Mother**
Elementary	21.46	23.62	59.51
Craft/trade	23.90	23.67	5.33
Clerk/sales	22.59	14.17	16.75
Operator/semi-skilled	23.96	29.33	5.21
Professional/manager	8.08	9.20	13.20
Unemployment rates	**Son**	**Father**	**Mother**
1993		25.87	33.04
2008	23.02		
2015	29.56		
N	**1 785**		

Source: Own calculations from the first four waves of NIDS. Attrition-corrected panel weights applied to members of the balanced panel. Unemployment rates from PALMS Version 3.1 are weighted using cross entropy weights.

Five different earnings variables were created for parents, and these correspond to five different imputation equations in the first stage of estimation. They use the following variables to impute parental earnings respectively: Education; education and race; education, race and occupation; education, race and province; education, race, occupation and province. Figure 4.A.1 in the appendix shows kernel densities of the log of earnings of fathers and mothers that were generated by the fifth imputation process. One possible way of assessing the quality of the imputation is to compare real earnings in the PSLSD to earnings that have been predicted using the variables in the fifth imputation procedure. Regressing age-adjusted log earnings on education, race, occupation and province produces an R-squared statistic of 0.63 for pseudo-fathers and 0.64 for pseudo-mothers in the PSLSD. A linear prediction of earnings from both of these regressions can be compared to the actual earnings variable in the data in order to get some idea of where the two may differ. The differences between predicted earnings and actual earnings in the PSLSD are very similar for pseudo-fathers and pseudo-mothers. Given the fact that we use a linear prediction, the means of predicted earnings and actual earnings are identical. The median of predicted earnings is slightly lower than that of actual earnings for both pseudo-fathers and pseudo-mothers. The most obvious difference between predicted and actual log earnings is the variance. For pseudo-fathers the predicted variance is 0.627 compared to the actual variance of 1.00, while for pseudo-mothers the corresponding numbers are 0.669 and 1.051, respectively. Although we do not estimate measures of earnings inequality using either predicted or actual earnings in this chapter, it is clear that inequality measures using predicted earnings will be lower than corresponding measures using the actual earnings data in the PSLSD.

These earnings are mapped against years of education and are shown in the education-earnings profiles in Figure 4.3. The real earnings of respondents in our balanced panel lie above those of their parents at every education level, and the same is true for father's earnings relative to mother's earnings. The convexity of the education-earnings profile of sons is evident, with a generally flat profile until the completion of secondary education, after which there are relatively higher returns to each year of postsecondary education.

Figure 4.3: Education-earnings profiles for the balanced panel and parents

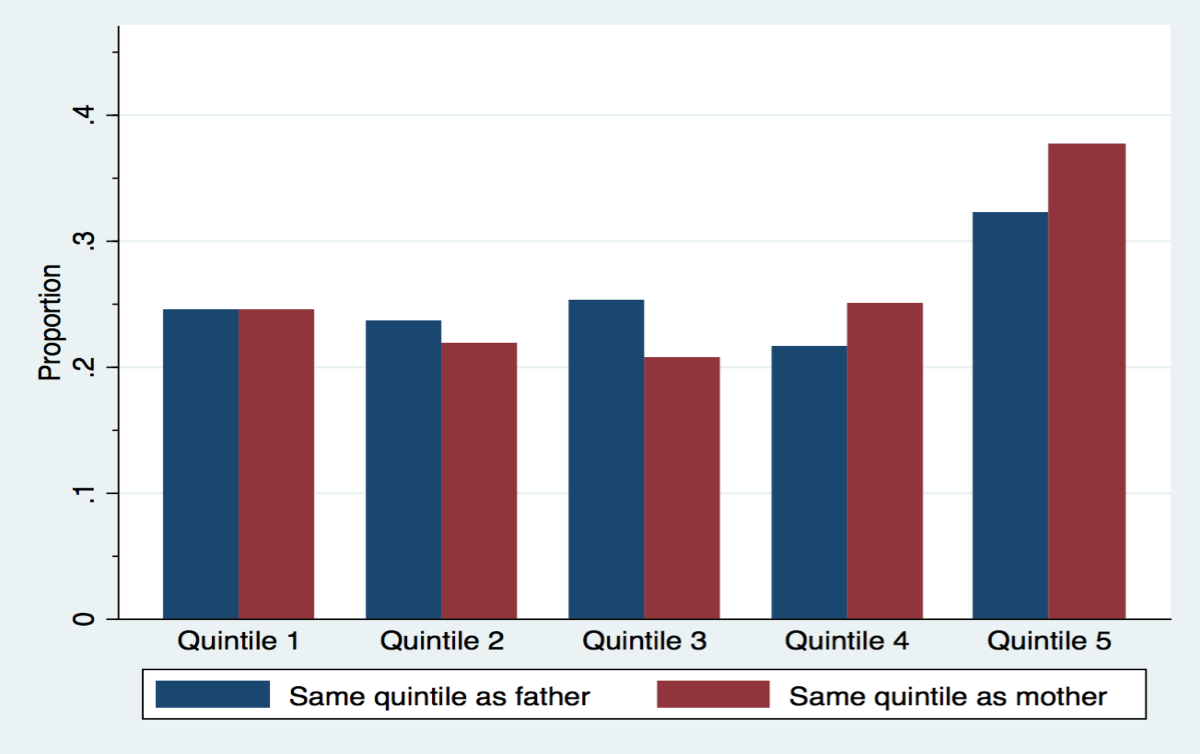


Source: Own calculations from the first four waves of NIDS. Attrition-corrected panel weights applied to members of the balanced panel.

We now turn our attention to the relative positions of parents and children in the distribution of earnings. Figure 4.4 below plots the probability that a son will be in the same earnings quintile as his parents.[16] Around one quarter of sons whose parents were in the bottom 20% of the earnings distribution are themselves in the bottom quintile. This proportion decreases to just under 20% for the middle quintile. Interestingly, sons whose fathers were in the 3rd earnings quintile are as likely to be in the bottom quintile or the top quintile themselves. There was relatively more downward mobility for sons whose parents were in the middle of the earnings distribution. Unsurprisingly, the highest probability of parent and child quintile matching is at the top of the earnings distribution. This top quintile shows a difference of about 5.5 percentage points between father and mothers, with child quintile matches of 32.3% and 37.7% respectively.

[16]The full transition matrices are presented in Table 4.B.1 in the appendix.

Figure 4.4: Unconditional probability of a son being in the same earnings quintile as his parents



Source: Own calculations from the first four waves of NIDS. Attrition-corrected panel weights applied to members of the balanced panel.

The general increase in educational attainment, as described in the first stylised fact in the introduction, is clear if we examine an educational transition matrix for parents and their children in the balanced panel. Table 4.2 shows the proportion of children in each education category, conditional on their parents being in a certain category. This reflects similar findings in Keswell et al. (2013) who use only the first wave of NIDS. The patterns for son's educational outcomes are similar whether we condition on the father's or the mother's highest attained level of education. Over a quarter of sons who had either a father or a mother with no education managed to complete at least a matric. There was very little downward educational mobility for sons whose parents had either a primary or an incomplete secondary education. The sample sizes for father and mothers with matric or postsecondary education are rather small, so the relatively large downward mobility for both of these categories should be interpreted with this in mind. It is important to note that though the increase in the general level of educational attainment has been large (particularly for the lower education categories) this presentation abstracts away from the quality of that increased education, though this is clearly an important part of understanding South Africa's labour market returns (Louw et al., 2007).

Table 4.2: Education transition matrices for parents and sons

		Son's education					
		None	**Primary**	**Inc. Sec.**	**Matric**	**Postsec.**	
Father's education	**None**	5.9	20.3	41.1	19.3	13.3	100
	Primary	0.3	9.8	49.8	18.5	21.5	100
	Inc. Sec.	0.0	7.5	41.1	28.0	23.5	100
	Matric	0.0	3.7	32.2	24.7	39.4	100
	Postsec.	0.0	0.3	21.6	26.6	51.5	100

		Son's education					
		None	**Primary**	**Inc. Sec.**	**Matric**	**Postsec.**	
Mother's education	**None**	7.3	20.3	41.9	18.5	11.9	100
	Primary	0.7	15.1	43.1	19.7	21.3	100
	Inc. Sec.	0.4	3.7	44.1	24.0	27.8	100
	Matric	0.0	1.0	42.7	29.4	26.9	100
	Postsec.	0.0	4.2	11.7	22.1	62.0	100

Source: Own calculations from the first four waves of NIDS. Attrition-corrected panel weights applied to members of the balanced panel.

Having shown that the increase in the level of education attained from generation to generation went hand in hand with an education-earnings profile that became more convex, we turn now to the estimation of the intergenerational earnings elasticity.

4.5 The intergenerational elasticity of earnings

Table 4.3 below presents the estimates of the intergenerational earnings elasticity between sons in the balanced panel and their fathers and mothers. Bootstrapped standard errors are presented along with the coefficients for each of the five columns, and all data are weighted using the attrition-corrected panel weights. Each numbered column represents a different imputation process for calculating parental earnings, and follows a similar sequence to Piraino (2015). In the first column the only variable used to predict parental earnings using the main and auxiliary datasets is the education of the parent. The number of variables used in the imputation process increases until column five, in which education, race, occupation, and province of residence in 1994 are used. In this table we have maintained the same sample for each estimation of the intergenerational earnings elasticity in order to ensure the comparability of our estimates, and to highlight the role that each additional variable plays in generating the intergenerational elasticity. If we did not apply this restriction then differences in sample sizes would arise based on the availability of parental information in the NIDS dataset.[17] Table 4.C.1 in the appendix shows that the unrestricted results are in line with the results in Table 4.3. For the remainder of this chapter we will restrict ourselves to the subsamples of sons who report all imputation variables for their parents - 1 389 fathers and 1 258 mothers, respectively. Though these sample sizes are slightly smaller than those in Table 4.C.1 in the appendix, they are nonetheless large enough to give us some confidence in the power of our calculations.

The elasticity relative to father's earnings ranges from 0.613 in the first column (education) to 0.678 in the third column (education and race). The elasticity is 0.659 if the province of residence of the father is added to education and race as an explanatory variable in the imputation equation. The fullest imputation, shown in column 5, reflects an elasticity of 0.627. Where comparable, these numbers are generally slightly lower than those reported in Piraino (2015), though it must be restated that the two studies use different sample members in their calculations and make different assumptions about weighting the data.

The degree of persistence relative to mother's earnings is also high, but differs in certain areas from the persistence relative to father's earnings. Imputing mother's earnings using only education generates an estimated elasticity that is about 4% higher than the corresponding figure for father's earnings. This differs slightly from the calculations in Piraino (2015), which find that the elasticity relative to mother's earnings is always lower than the elasticity relative to father's earnings. In fact, we find that the elasticity relative to mother's earnings is higher for all imputation procedures except for when education, race and occupation are used jointly. The difference is reinforced if earnings are imputed using all four of the available variables - from 0.627 for fathers to 0.650 for mothers.

[17]For example, more sons provide information about parental education than parental occupation.

Table 4.3: Intergenerational earnings elasticities for different imputation procedures

	Variables used to construct parental earnings				
	1 Education	**2** Education, race	**3** Education, race, occupation	**4** Education, race, province	**5** Education, race, occupation, province
			Father's earnings		
Elasticity	0.613	0.678	0.674	0.659	0.627
	(0.159)	(0.186)	(0.188)	(0.187)	(0.187)
N	1,389	1,389	1,389	1,389	1,389
			Mother's earnings		
Elasticity	0.639	0.693	0.592	0.754	0.650
	(0.184)	(0.176)	(0.168)	(0.184)	(0.170)
N	1,258	1,258	1,258	1,258	1,258

Source: Own calculations from the first four waves of NIDS. Attrition-corrected panel weights applied to members of the balanced panel. Bootstrapped standard errors in parentheses.

4.5.1 Accounting for selection into employment

The matter of selection bias is something that always underlies estimates of intergenerational mobility. Indeed, the possible bias arising from not modelling female labour force participation decisions is a major reason for why daughters' earnings are usually not reported in these kinds of studies. Another bias already mentioned is the selection bias that may arise from restricting the analysis to children who co-reside with their parents. This is dealt with in this chapter by the use of the TSTSLS estimator. There is, however, another selection issue that is often ignored in the international literature that we may want to consider, and that is selection bias arising from who finds a job and who does not. We only observe the earnings of those who are employed, and it may be that both labour market participation decisions and finding employment are not random. This is a particularly pertinent issue in South Africa, given that unemployment rates are high in general, and are very high for youth in particular (Ranchhod and Finn, 2016). The structure of the South African labour market and the relatively high demand for high-skilled workers means that it is possible that we calculate a biased elasticity when we do not take selection into employment into account. It is possible that those potential workers with parents whose earnings were low are less likely to find employment themselves. In a counterfactual world in which we observe earnings for all our respondents (rather than only for those who are employed), we may find that correcting for selection matters in the measurement of the correlation between parental and child earnings. However, applying the correction only to sons ignores the fact that the pseudo-parents in the 1993 dataset faced similarly high unemployment rates, and that the coefficients extracted from the first stage imputation may be biased as well. We are therefore faced with an estimating equation that requires two corrections – one in the first stage when the parental earnings variable is imputed, and one in the second stage when the intergenerational earnings elasticity is calculated.

Essentially the problem is as follows. What we would like is an unbiased estimate of the relationship between parental earnings and son's earnings. In a world without unemployment, we would have a full set of earnings for parents and sons. However, in South Africa, with unemployment rates being so consistently high (generally more than 25%), the subsample that we see employed may not be random, and therefore the earnings associated with this subsample may be biased. As noted in Vella (1998), if the subsample of the employed is a random selection of the population, then there is no selection bias problem because the average observed and unobserved characteristics of the subsample are the same (or in fact not significantly different) as those of the population. Assume now that there are differences between the employed and the unemployed, so that the subsample of the employed is no longer a random sample of the population. If the differences between the employed and unemployed are generated only through observable characteristics, then one may arrive at an unbiased wage by controlling for these observables in the wage equation. In other words, if there are observables that are correlated with both the decision to work, and with wages themselves, then selection bias is not a problem if those characteristics are controlled for. Finally, assume that the unobservables determining the decision to work and the unobservables determining wages are correlated. In this case measured wages will be biased because of sample selection, even if observable characteristics are controlled for in the wage equation.

The correction for this selection bias is implemented by deriving a fully parametric expression for the expected value of wages, via the calculation of the inverse Mills ratio, which is conditional both on observable characteristics and on selection into employment. What this gives us is, in some sense, a counterfactual of what earnings would be if the wages of the unemployed could be observed. This can also be thought of as 'potential' earnings. Estimating the IGE with a double correction for selection bias can therefore also be thought of as a calculation of the association of potential earnings between parents and children.

In this chapter we correct for possible selection bias into employment for both parents and children by using a two-stage model of the type that was proposed for modelling selection into employment by Gronau (1974) and Heckman (1974), and has been used in the intergenerational mobility literature by Ermisch et al. (2006) among others.

In the first stage we use a probit to model whether a respondent is employed (and therefore earning a wage) or not. Variables included in this selection equation but not in the outcome equation are a dummy for the presence of dependent children in the household, marital status, age, and parental earnings. The first two of these variables are included so that the model is identified by exclusion restrictions, rather than by the non-linearity of the first stage.[18] We generate the correction term (the inverse Mills ratio) which can be thought of as capturing the 'surprise' of observing an individual who is employed and earning. In other words, the residuals from the first stage are captured by the inverse Mills ratio. For example, a respondent who has a job but also has a low level of education will have a larger residual, and therefore a higher inverse Mills ratio, than a respondent with postsecondary education who is employed. Our results can therefore be thought of in somewhat clumsy terms as being derived from a two sample, two stage, twice corrected least squares (TSTSTCLS) estimator.

Correcting for selection into employment yields elasticities that are higher than the 'naïve' estimation for son's earnings relative to fathers and mothers. Employment selection biases our uncorrected elasticity downwards for fathers – the corrected elasticity is 0.678 compared to an uncorrected elasticity of 0.627. For mothers the bias is in the same direction and of an even greater magnitude – a corrected elasticity of 0.718 compared to an uncorrected elasticity of 0.650. The full results of this double correction are presented in Table 4.4 below. This is our preferred set of results in general, with the elasticities in column 5 being the preferred point

[18] Full results are available from the author.

estimate in particular. Once again we restrict ourselves to the subsample of sons who report full information on parental background. The unrestricted sample estimates can be found in Table 4.D.1 in the appendix.

Table 4.4: Intergenerational earnings elasticities for different imputation procedures with a double Heckman correction

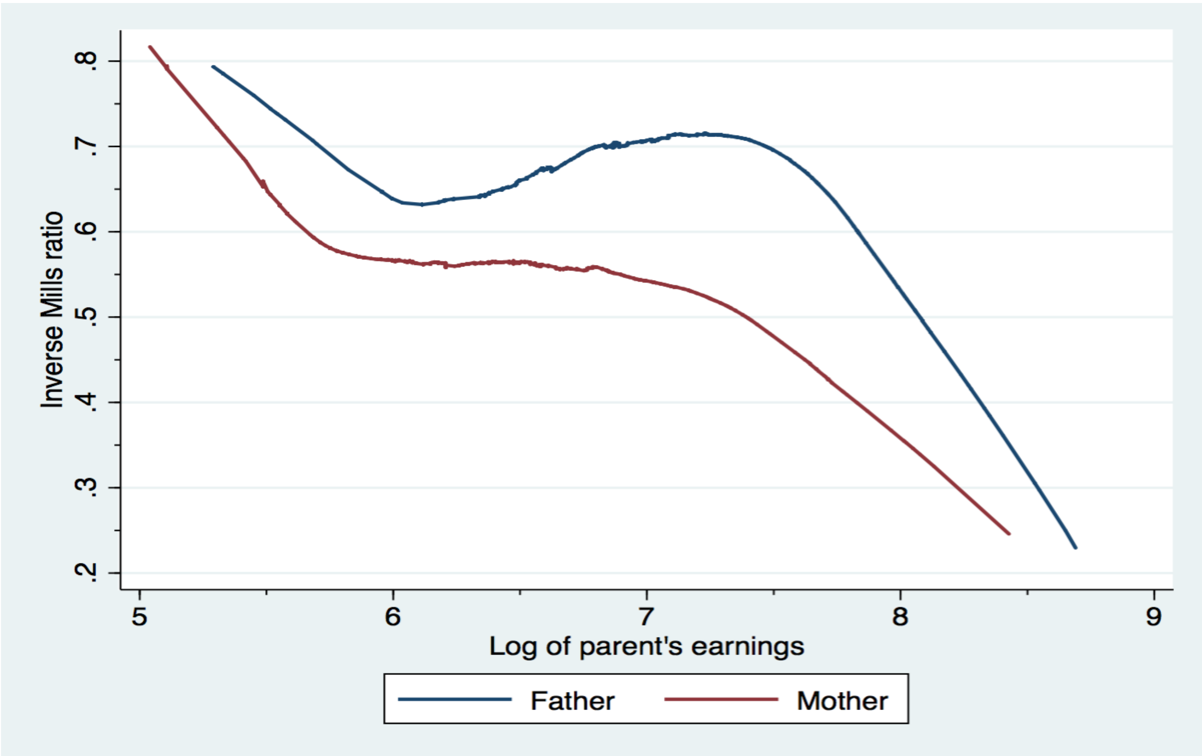
	Variables used to construct parental earnings				
	1 Education	**2** Education, race	**3** Education, race, occupation	**4** Education, race, province	**5** Education, race, occupation, province
			Father's earnings		
Elasticity	0.612	0.718	0.697	0.723	0.678
	(0.214)	(0.234)	(0.220)	(0.204)	(0.215)
N	1,389	1,389	1,389	1,389	1,389
			Mother's earnings		
Elasticity	0.659	0.739	0.650	0.825	0.718
	(0.225)	(0.247)	(0.221)	(0.214)	(0.220)
N	1,258	1,258	1,258	1,258	1,258

Source: Own calculations from the first four waves of NIDS. Attrition-corrected panel weights applied to members of the balanced panel. Bootstrapped standard errors in parentheses.

We can investigate our intuition that children with low-earning parents are less likely to find a job themselves by plotting the inverse Mills ratio over the range of parental income. In Figure 4.5 the inverse Mills ratios are presented for fathers and mothers over their respective earnings ranges. The higher the line, the more 'surprised' we are to see an individual in a wage-earning job, given parental earnings. The figure accords with our intuition in that the ratio decreases as we move rightward across the parental earnings distributions. Those with parents who earned relatively higher salaries are more likely to be employed than those with parents who earned at the lower end of the distribution. The inverse Mills ratio for the log of mother's earnings drops sharply, then flattens out, and then drops again as we move rightward along the distribution. The pattern for fathers is slightly different as the ratio first drops, then rises, and then drops off sharply. This suggests that the 'surprise' at seeing a son in employment, conditional on his father's earnings, does not decrease monotonically across the distribution of earnings.

The hump in the son's inverse Mills ratio relative to father's earnings is driven primarily by the role of father's occupational category in the imputation of earnings. In particular, the shape of this line comes from fathers who were employed in elementary occupations in 1993. These made up almost a quarter of the fathers in our sample. The earnings for this category are concentrated at the bottom of the distribution, though it does have quite a long right tail. The sons of fathers who were employed in elementary occupations whose earnings were at the bottom of the distribution, were very unlikely to be employed themselves. The same is true for the sons of fathers who were employed in elementary occupations, but who earned towards the middle of the distribution (between $ln(6.5)$ and $ln(7.5)$), but less so for the sons of fathers who earned between $ln(6)$ and $ln(6.5)$.

Figure 4.5: Inverse Mills ratio over the distribution of parental earnings



Source: Own calculations from the first four waves of NIDS. Attrition-corrected panel weights applied to members of the balanced panel.

The intergenerational elasticities that we report are high by international standards, but focusing on a single number may hide underlying patterns. The heteroskedasticity present in the sample[19] means that quantile regression analysis is a potentially useful tool in evaluating the

[19]The White test for heteroskedasticity rejects the null of constant variance for all specifications of the regres-

joint distribution of parental and child earnings. To this end we run quantile regressions from the 5th to the 95th percentile, increasing in intervals of 5. As described in Buchinsky (1998), we estimate the coefficient vector β as the solution to the following:

$$\min_{\beta(\theta)} \left\{ \sum_{i:y_i \geq x_i\beta(\theta)} \theta|y_i - x_i\beta(\theta)| + \sum_{i:y_i < x_i\beta(\theta)} (1-\theta)|y_i - x_i\beta(\theta)| \right\}$$

where y_i is son's earnings, x_i is the earnings of either the father or the mother, and θ is the quantile being estimated.

Quantile regression analyses of intergenerational mobility in low-inequality countries have found that the correlation between parental and child income falls over the distribution of earnings. For example, Bratberg et al. (2007) use Norwegian data and find a monotonic decline in the intergenerational elasticity for men, and a decreasing but non-monotonic fall for women in Norway from the 5th to the 95th percentile, showing that earnings persistence is far higher at the bottom of the earnings distribution than at the top.

Studies using data from the US consistently find that persistence is highest at the bottom of the earnings distribution, but disagree as to what happens to the correlation as earnings increase. Eide and Showalter (1999), using a rather small sample of American father and son pairs, find a decreasing pattern with a slight upturn at the very top of the earnings distribution. A relatively higher correlation between parental and child earnings at the bottom of the child's earning distribution in the US is also found by Lee et al. (2009) A slightly different pattern emerges in a recent paper by Palomino et al. (2014) who use a much larger sample of US data and find what they refer to as a 'U' shape, indicating that persistence is highest at the bottom of the earnings distribution, but that there is an upturn at the top of the distribution as well.[20] It is likely that high-inequality societies produce a U-shaped relationship between the intergenerational elasticity and earnings. High cross-sectional inequality is stable over time if there is high persistence between both low-earning parents and their children, as well as high-earning parents and their children. Given how high and persistent inequality in South Africa has been over the last two decades, we might expect to see a turning point in the elasticity-earnings

sion.

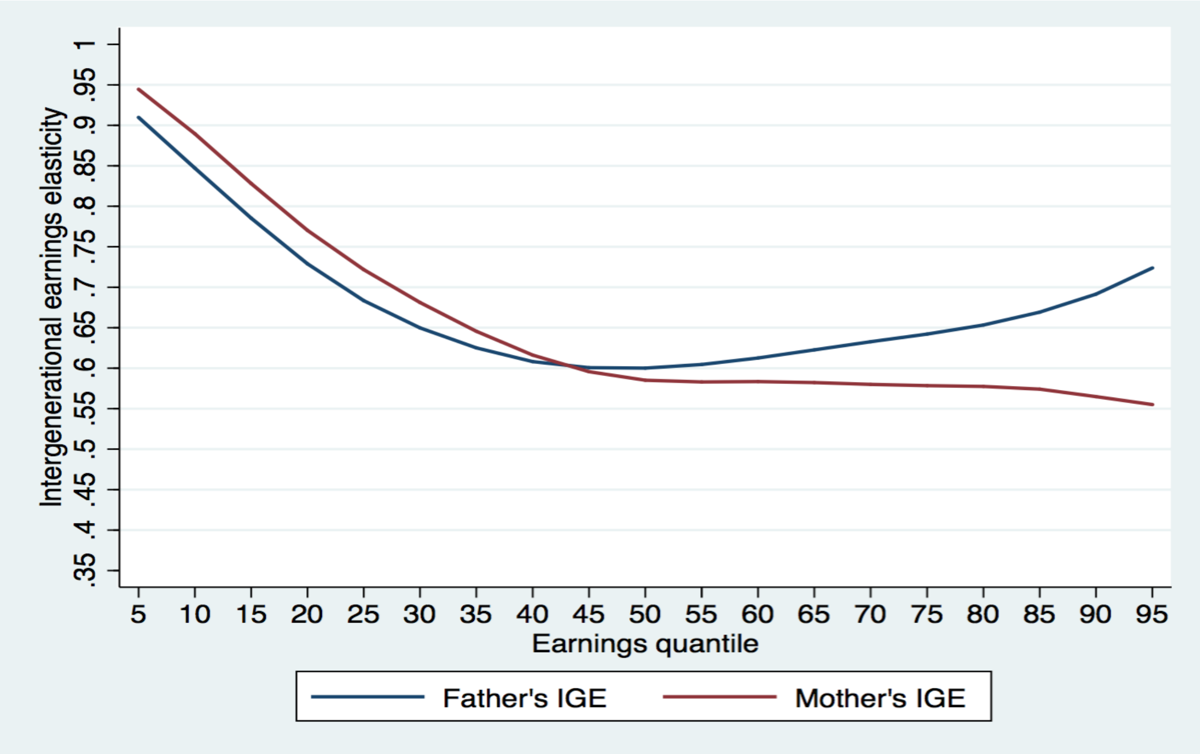
[20]The Palomino et al. (2014) paper finds that the turning point occurs around the 70th percentile.

relationship.

Figure 4.6, below, plots smoothed versions of the double corrected intergenerational earnings elasticities for South Africa between the 5^{th} and 95^{th} percentiles of the earnings distribution. It is clear that relying only on the conditional mean hides a great deal about the pattern of persistence in the country. The intergenerational elasticity is highest at the bottom of the distribution, and this accords with the international evidence for both developed and developing countries. What is different about the South African case is the fact that the persistence is so high in this part of the distribution - over 0.9 for both mothers and fathers at the lowest end. This shows that the low-earning sons have a far higher correlation with their parents' wages than high-earning sons do with theirs. There is an interesting difference in the shapes of parental elasticities. The strength of the association between son's earnings and mother's earnings decreases monotonically as we move rightwards across the distribution of earnings. For father's earnings, however, a turning point is reached at around the 40th percentile, after which there is an increase to about 0.73 at the top of the distribution.[21]

[21] Quantile graphs for all the different imputed versions of father's and mother's earnings are available from the authors on request.

Figure 4.6: Intergenerational earnings elasticity over earnings quantiles



Source: Own calculations from the first four waves of NIDS. Attrition-corrected panel weights applied to members of the balanced panel.

4.6 The role of education in shaping intergenerational mobility

We build on the previous section by investigating the role that education plays in shaping intergenerational mobility in South Africa. The ideal set of data for getting precise estimates of various transmission mechanisms would include child's ability, parent's ability and school quality. Although we are able to make use of a rich dataset, we do not have all of these variables available and so we must find more indirect ways of getting at the relationship between education and intergenerational mobility.

One way of doing this is to follow Palomino et al. (2014) by measuring the strength of the association between child's education and the intergenerational elasticity of earnings by quantiles by including the child's level of education as an additional regressor in the canonical regression in equation 4.2. We can think about the effect that including child's education would have on the elasticity in the same way that we think about omitted variables in OLS regressions. Retaining the representation of parental earnings as Y_{it}^p and using Edu_i^c as the variable indicating child's education[22] (which is omitted from equation 4.2), we can represent the elasticity as:

$$plim\hat{\beta}^{OLS} = \beta + \beta_{Educ_i^c} \times \frac{cov(Y_{it}^p, Edu_i^c)}{\sigma^2_{Y_{it}^p}}$$

This equation can be used to interpret what happens to the intergenerational elasticity when we add a control for child's education into the estimating equation. If there is a strong positive correlation between parental earnings and child's education, then the elasticity as estimated in equation 4.2 will be higher if education is not controlled for. On the other hand, if there is a zero correlation between parental earnings and the child's education then there will be no change in the estimated elasticity once a control for education is included. This is true even in the presence of the relationship between education and earnings for the child.

Including child's education in the estimation of the intergenerational elasticity of earnings reduces the double corrected elasticity at the mean by 41% and 39% relative to father's and mother's earnings respectively. The relationship between education and the intergenerational elasticity changes along the distribution of earnings, however, and this is shown in Figure 4.7 below.[23] The estimation procedure underlying the figure is the same as it was for Figure 4.6. Intergenerational elasticities are estimated for parental earnings from the 5$^{\text{th}}$ to the 95$^{\text{th}}$ percentile. The vertical axis shows the percentage difference in the intergenerational elasticity for a regression that includes child's education versus one that does not. The larger the negative difference between the elasticities in a particular quantile, the higher the positive correlation between education and parental earnings in that quantile.

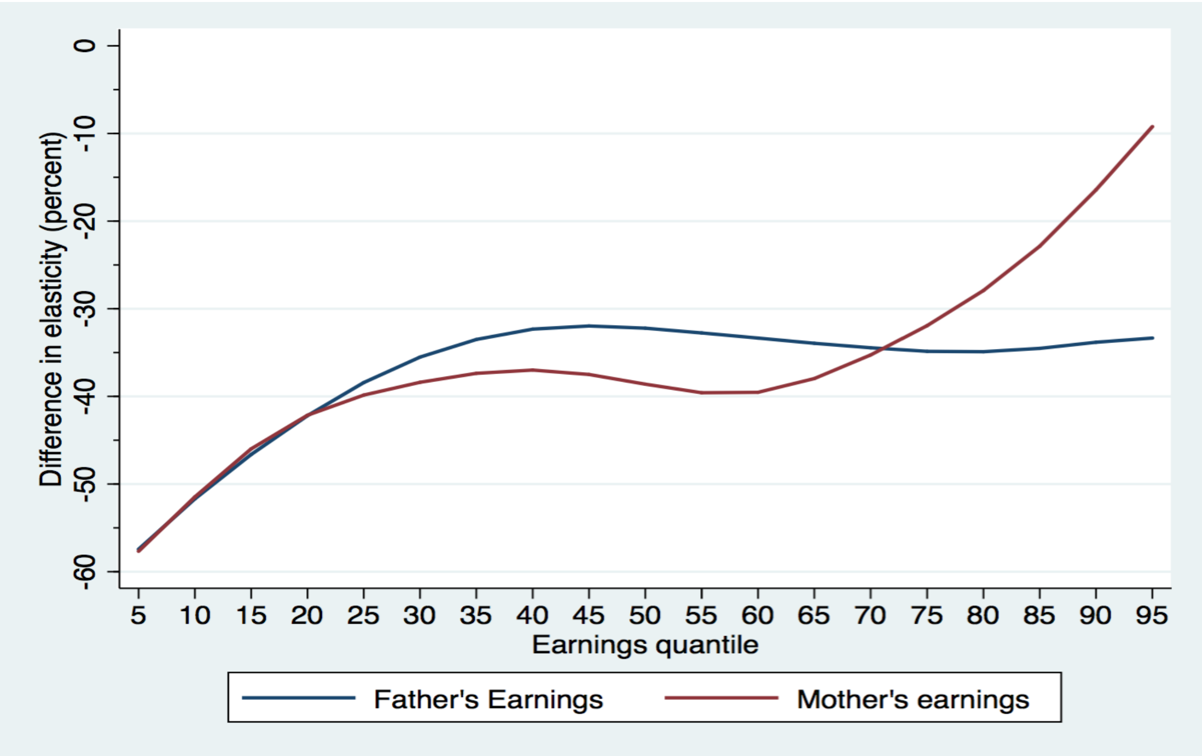
The relationship between educational attainment and parental earnings follows a different shape over the distribution of child's earnings depending on whether we look at mother's earnings or father's earnings. Including the child's education as an additional control has the largest negative effect at the bottom of the distribution for both parents - reducing the coefficient by

[22] Though there are a few exceptions in the data, the level of education attained by each child is time-invariant across the four waves. Those respondents whose education status changes are generally those who move from matric to postsecondary. For sons whose education changes over the four waves we use the level of education reported in the fourth wave.

[23] As in Figure 4.6, this figure presents estimates that are smoothed using a LOWESS procedure.

close to 50%. For mothers this effect is generally decreasing as we move up the earnings quantiles, and is almost negligible at the top of the distribution. There is a low correlation between mother's earnings and child's education at the top end of the distribution, and this corresponds to the part of the earnings distribution with the lowest level of intergenerational elasticity. One of the insights of this figure is that the further up the earnings distribution we travel, the less important educational attainment is in explaining the level of mobility between parental and child earnings. For fathers the pattern is slightly different. The correlation between child's education and parent's earnings is strongest at the bottom of the distribution, and the strength of this relationship decreases steadily until the 35th percentile of earnings. Thereafter it remains relatively flat, with roughly the same correlations at the 35th and 95th percentiles.

Figure 4.7: Difference in intergenerational elasticity when controlling for education, over parental earnings quantiles



Source: Own calculations from the first four waves of NIDS. Attrition-corrected panel weights applied to members of the balanced panel.

Figure 4.D.1 in the appendix shows the graph for African respondents only. The effect of education is even larger at the bottom of the African wage distribution, and the line for mothers crosses the line for fathers at a lower percentile (the 35th) compared to the lines for all

respondents (the 70$^{\text{th}}$). In Figure 4.7 the role of education at the top of the wage distribution is muted for mothers (-10%), but still fairly large for fathers (-32%). In the African subsample, the effect for fathers is similar to what it was in Figure 4.7 (30%), but the effect for mothers at -25% is much larger in absolute terms than in the corresponding figure for all respondents.

Another way of extracting the role of education in determining intergenerational mobility is to decompose the intergenerational elasticity into a component that is due to education and a component that is due to parental earnings. Blanden and Macmillan (2014), referencing an earlier model by Blanden et al. (2007),[24] break the estimation of the intergenerational elasticity into two stages. This allows us to look at the relationship between parental characteristics, child characteristics, and the labour market returns to these characteristics when the child is working. Essentially, this is a standard path model decomposition in which the direct and indirect effects of education on earnings are separated. These decompositions reflect some of the earlier empirical work (for example see Conlisk (1971)) in which the estimating equation had a very similar structure.

In the first stage we regress the child's level of educational attainment on the log of parental income. In the second stage we regress the child's income on his education and parental income - this is the same estimating equation underlying the previous figure. The two equations are:

$$edu_i^c = \hat{\alpha}_2 + \gamma Y_i{}^p + \hat{e}_i \tag{4.7}$$

and

$$Y_i^c = \hat{\alpha}_2 + \hat{\rho} edu_i^c + \hat{\delta} Y_i^p + \hat{u}_i \tag{4.8}$$

Taken together, these two equations decompose the intergenerational elasticity into the contribution of education inequality ($\hat{\gamma}$), the contribution of the returns to education ($\hat{\rho}$), and the influence of parental income on child's income (controlling for child's education). Blanden et al. (2007) show how the intergenerational elasticity can be written as:

[24] Originally this was done in order to separate out the relative importance of cognitive versus non-cognitive skills in the association of parental and child earnings.

$$\hat{\beta} = \hat{\gamma}\hat{\rho} + \hat{\delta} \tag{4.9}$$

According to this formulation the contribution share of education variables[25] to the overall intergenerational elasticity at the mean is close to 40% for father's earnings and 43% for mother's earnings.

In our final decomposition we turn to the question of the intergenerational transmission of occupational skill, and how this shapes the intergenerational earnings elasticity compared to the role that education plays. We follow Keswell et al. (2013)[26] and use the occupational codes in the NIDS dataset as proxies for the skill level of each respondent and his parents. The skill level is derived using the same method and variable as Keswell et al. (2013) who map the SASCO occupational codes to skill levels. This, in turn, follows Bergman and Joye (2005) who use the occupational codes to classify work according to a) which tasks and duties are related to an occupation, and b) which relevant skills are necessary for required for fulfilling the requirements of each particular occupation. This enables us to transform the SASCO coded occupational variable into a hierarchical occupational skill variable for use in the decompositions.

The original decomposition of the intergenerational elasticity of earnings into education and skill components can be found in Bowles and Gintis (2002), and it was quickly adopted in the economics literature (two recent examples are Lefranc and Trannoy (2005) and Cervini-Plá (2013)). We use Lefranc and Trannoy's notation in explaining this decomposition. It is important to note that this is not to be interpreted as a 'causal' decomposition in the traditional sense, but rather as an attempt to extract the relative importance in the correlations between parental and child education versus occupation in generating the intergenerational earnings elasticity. It is also important to note that the equations used in this decomposition assume a linear education-earnings profile and therefore homogeneous returns to each additional year of education. The imposition of a linear structure on what has been shown to be a convex relationship means that the effect of education on the intergenerational transmission of earnings is likely to be overestimated for those respondents with low levels of schooling, and underesti-

[25] We do not separate out educational attainment and returns to education at this stage.

[26] The authors link educational opportunity to the distribution of steady state occupations in South Africa using the first wave of NIDS data.

mated for those respondents with postsecondary education. Performing the decomposition for different quantiles, and using more flexible functional forms of the education variable is left for future research.

Let us assume that for $g = c, p$, parental and child income may be expressed as:

$$Y_i^g = Edu_i^g \gamma_e^g + Skill_i^g \gamma_s^g + \nu_i^g \tag{4.10}$$

The TSTSLS estimate of β derived from this relationship is:

$$\beta = \frac{cov(Y_i^c, Edu_i^p \gamma_e^p + Skill_i^p \gamma_s^p)}{V(Edu_i^p \gamma_e^p + Skill_i^p \gamma_s^p)} \tag{4.11}$$

We expand β using equation 4.11 so that:

$$\begin{aligned} \beta = & \frac{1}{V(Edu_i^p \gamma_e^p + Skill_i^p \gamma_s^p)} \\ & \times [\gamma_e^c cov(Edu_i^c, Edu_i^p)\gamma_e^p + \gamma_s^c cov(Skill_i^c, Edu_i^p)\gamma_e^p \\ & + \gamma_e^c cov(Edu_i^c, Skill_i^p)\gamma_s^p + \gamma_s^c cov(Skill_i^c, Skill_i^p)\gamma_s^p \\ & + cov(\nu_i^c, Edu_i^p)\gamma_e^p + cov(\nu_i^c, Skill_i^p)\gamma_s^p] \end{aligned} \tag{4.12}$$

β has been decomposed into six terms comprising the covariances of the child and parental education and occupational skill, and the covariance of the child's earnings residual and parental education and skill. These are multiplied by the relevant coefficients from equation 4.10.

In Table 4.5 each row represents the contribution shares of each term in the decomposition to the overall intergenerational earnings elasticity. The relationship between father's education and son's education accounts for 37.6% of the intergenerational elasticity. The corresponding share for the mother-son elasticity is slightly lower at 34%. The intergenerational correlation of occupational skill is less important in determining β than the intergenerational correlation of education - 9% for both fathers and sons and mothers and sons. These contribute approximately the same as the 'cross' correlations of parental education and the child's occupational skill. The correlation between parental skill and child's education is the smallest contributor to the IGE in both the father to son and mother to son panels. The relatively large contribution share of education compared to occupational skill in determining the IGE is heightened when we consider the African subsample only, as is done in the second column of results in the table. The contribution share of the correlation between parental and child education is 44% for fathers and sons, and almost 50% for mothers and sons. The correlations between parental education and the unexplained (residual) part of the son's wage equation are also large, and indeed are far larger than the corresponding share from the correlation of parental skill with the residual from the son's wage equation. The contribution of the correlation between father's education and son's skill is far larger than the corresponding share for mothers, and the relationship between parental skill and son's education is muted in both panels.

It therefore appears that the joint impact of parental education on son's education and occupational position is far larger than the joint impact of parental occupational skill through the same channels. This is in contrast to studies in OECD countries by Cervini-Plá (2013), Lefranc and Trannoy (2005) and Österbacka (2001) who find that parental social position, rather than parental education, is the most important determinant of intergenerational mobility.[27]

[27]Here, social position refers to the schema suggested by Erikson and Goldthorpe (1992) which consists of the following seven classes: higher-grade professionals, lower-grade professionals, routine non-manual employees (administration and commerce), routine non-manual employees (sales and service), lower-grade technicians, skilled manual workers, semi- and unskilled manual workers.

Table 4.5: Contribution shares to intergenerational elasticity - education versus skills

Fathers and sons		
	All	**African**
Edu. father, edu. son	37.56	44.22
Skill father, skill son	8.89	4.92
Edu. father, skill son	8.62	14.34
Skill father, edu. son	2.32	2.59
Edu. father, resid. son	40.16	36.44
Skill father, resid. son	2.45	-2.52
Mothers and sons		
	All	**African**
Edu. mother, edu. son	34.22	49.87
Skill mother, skill son	8.94	7.00
Edu. mother, skill son	7.91	2.98
Skill mother, edu. son	3.51	0.86
Edu. mother, resid. son	36.24	37.19
Skill mother, resid. son	9.16	2.10

Source: Own calculations from the first four waves of NIDS. Attrition-corrected panel weights applied to members of the balanced panel.

4.7 Discussion

It is abundantly clear that there is not a level playing field in South Africa in terms of equality of opportunity. This is manifest in the differential probabilities of finding work based on parental earnings, as well as the high correlations of intergenerational earnings at the bottom and the top of the distributions, as shown in this chapter. One of the key questions is why the children of low-earning parents have been unable to translate greater educational attainment into better labour market outcomes.

Intergenerational mobility is a complex process which is generated by individual decisions, family and social norms, and public policies. Studying intergenerational earnings mobility is one way of thinking about equality of opportunity, but it does not leave one with a comprehensive understanding of the full process. However, an example based on our results can highlight just how stark this immobility is.

If we assume that the long-run log earnings of fathers and sons are of equal variance, and are distributed bivariate normal, then we can derive some back-of-the-envelope calculations about the probabilities of shifts in the relative distribution. For example, for our estimated intergenerational earnings elasticity of 0.678, the probability that a son is in the top half of the earnings distribution if his father was in the 5^{th} percentile of the earnings distribution, is just over 5%. If the intergenerational earnings elasticity were zero, that probability would be 50%. Alternatively, a son whose father earned at the 90^{th} percentile of the earnings distribution, has about a 28% chance of being in the top 10% of the earnings distribution himself, and has over a 60% probability of being in the top quarter of the earnings distribution.

The beginning of this chapter highlighted three features of post-apartheid South African society. These were the rapid expansion of educational attainment, the increasing returns to postsecondary education, and the stubbornly high level of economic inequality. Although the average level of education attained by South Africans increased rapidly, the number of South Africans enjoying the high returns to tertiary education remains relatively low. It would seem that the education South African children receive at primary and secondary level - both in terms of content and quality - is simply not matching up sufficiently to what the current labour market is demanding. It has become something of a truism in the South African discourse to say this, but the only way that this can change is with a concentrated improvement in educational quality and outcomes at the primary and secondary levels. As discussed earlier in this chapter, the question of school quality plays an important role in generating differential dropout rates for different groups of the population. These adverse effects are disproportionately born by pupils attending historically black schools. Stopping pupils from falling behind in the first place is a crucial part of addressing dropout and its resulting barrier to tertiary education.

Policy interventions and their ethical justification may depend on one's assumptions about the equal or unequal distribution of individual abilities. The South African evidence suggests that the structural nature of immobility and inequality of opportunity has less to do with individual characteristics and more to do with the inheritance of advantage and disadvantage.

What should the role of public policy be? Of course, there are some factors which determine the level of intergenerational mobility that public policy can only impact marginally upon. Social norms and the extent of social networks are two examples. Policy makers can prioritise helping the poor escape poverty, curtailing the relative advantage of the wealthy, or some combination of the two. This chapter has shown that the relative lack of intergenerational mobility is being driven by both factors: a great many South Africans are trapped in low earnings and household poverty dynamically,[28] while there is very little mobility at the top of the earnings and income distributions. One clear role for social policy, given the findings of this chapter, is to reform and improve the public education system of the country. Greater access to tertiary education for given entry requirements cannot simply be imposed - it has to start with improvements at primary and secondary levels. This is especially important because of the central role that education plays in determining the intergenerational correlation of earnings (see Figure 4.7 and Table 4.5, for example).

Another option that could have implications for intergenerational mobility is for policymakers to intervene directly in the labour market. The most prominent recent example of this kind of approach (though not instituted with concerns about intergenerational mobility directly in mind) is the youth employment tax incentive. This intervention is theoretically appealing, as it aims to reduce the cost to employers of hiring youth for new positions, with the additional benefit of increasing the labour market experience for the youth. However, in practice, early results using quarterly labour force survey data suggest that the policy has not had a significant impact on youth unemployment rates in the short term (Ranchhod and Finn, 2015, 2016).

Another recent intervention aimed directly at the bottom of the earnings distribution is the announcement of a national minimum wage (National Treasury, 2016). Although any evaluation of such an ambitious policy intervention must take general equilibrium effects into account, it will be interesting to see what the effects of raising the wage floor has on mobility in the country in the medium-to-long term.

4.8 Conclusion

One of the social questions that sparked this study is why earnings inequality in South Africa has remained so high from one generation to the next in the face of increasing educational attainment. The dynamics of intergenerational earnings imply that the higher the intergenera-

[28] This links to the findings in the chapter on poverty dynamics.

tional elasticity, the longer it will take for a convergence in earnings in society to take place. As a first step to uncovering some of the underlying intergenerational patterns we followed the methodology outlined in Piraino (2015) and calculated the intergenerational earnings elasticity for a balanced panel of South African males. We corrected for two kinds of bias in the estimation of the intergenerational elasticity. The first - co-resident selection - was mitigated through the use of a TSTSLS estimator. The second - selection into employment in a high-unemployment society - was corrected through the use of a Heckman two-step procedure.

We found that although the intergenerational elasticity of earnings is very high (implying low mobility) it varies markedly over the distribution of earnings. The degree of association between parental and child earnings changes along the distribution of earnings. It is highest at the bottom of the distribution and then falls until the middle of the distribution. For mothers this trend continues, and the association is weakest at the top of the distribution. For fathers, however, there is a turning point, and the correlation rises until reaching approximately 0.73 for the 95th percentile.

We then tried to reconcile the high association between parental and child earnings with the rise in educational attainment in the country over the last two decades. Other studies have found that although schooling attainment has increased in South Africa, the returns to education remain convex. This implies that even if the younger generation is more educated than the older generation, there will not necessarily be a proportional increase in earnings. We found that the correlation between education and the intergenerational persistence of earnings is highest at the bottom of the earnings distribution, and that the pattern of this correlation over the first half of the distribution is the same whether father's or mother's earnings are the focus. Thereafter the correlation between education and mother's earnings decreases steadily, while the correlation between education and father's earnings remains roughly the same. Finally, we made use of two different decompositions of the intergenerational earnings elasticity, and found that education accounts for around 40% of the elasticity, and that education plays a greater role in understanding earnings persistence than does occupational skill.

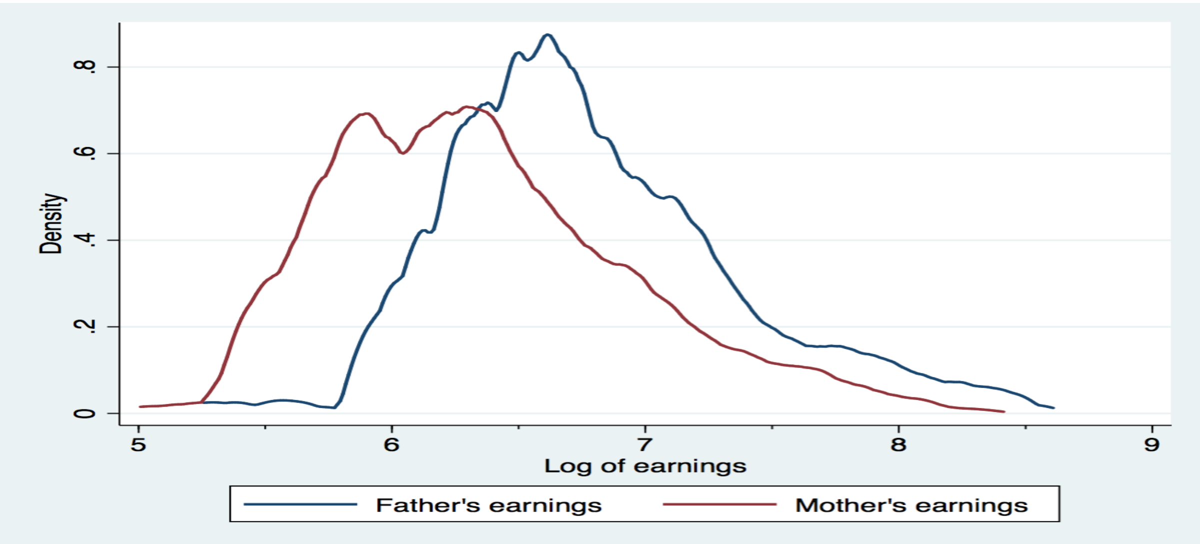
One issue that we did not touch upon is the quality of education in South Africa. This refers to both the average quality and the variance in quality across educational institutions.

Though there has been steady growth in the access to education in South Africa, it is debatable whether there has been a concomitant rise in the quality of that education. Given the richness of the NIDS dataset and the possibility of linking respondents to administrative data, uncovering the roles of the education quality versus quantity in shaping intergenerational earnings and persistent inequality is something that may be possible in the future.

Appendix

4.A Distributions of parental earnings

Figure 4.A.1: Kernel densities for imputed parental earnings (specification 5)



Source: Own calculations from the first four waves of NIDS and the PSLSD. Attrition-corrected panel weights applied to members of the balanced panel.

4.B Earnings transition matrices

Table 4.B.1: Earnings transition matrices

		Son quintiles					
		1	**2**	**3**	**4**	**5**	
Father quintiles	**1**	24.5	27.6	18.8	17.2	11.9	100
	2	25.1	23.7	19.5	23.0	8.6	100
	3	22.5	18.1	25.3	12.9	21.3	100
	4	18.3	18.2	17.1	21.6	24.8	100
	5	9.7	11.8	19.4	26.8	32.3	100
		Son quintiles					
		1	**2**	**3**	**4**	**5**	
Mother quintiles	**1**	24.6	25.1	22.6	16.9	10.9	100
	2	23.5	21.9	16.8	16.4	21.4	100
	3	20.6	21.6	20.8	18.8	18.2	100
	4	16.2	14.3	27.4	25.1	17.0	100
	5	6.9	12.9	14.9	27.6	37.7	100

Source: Own calculations from the first four waves of NIDS. Attrition-corrected panel weights applied to members of the balanced panel.

4.C Elasticities for different imputations and different subsamples

Table 4.C.1: Intergenerational earnings elasticities for different imputation procedures for different subsamples

	Variables used to construct parental earnings				
	1 Education	**2** Education, race	**3** Education, race, occupation	**4** Education, race, province	**5** Education, race, occupation, province
	Father's earnings				
Elasticity	0.634	0.706	0.682	0.680	0.627
	(0.166)	(0.204)	(0.205)	(0.194)	(0.187)
N	1,782	1,782	1,397	1,774	1,389
	Mother's earnings				
Elasticity	0.615	0.689	0.601	0.723	0.650
	(0.197)	(0.185)	(0.158)	(0.181)	(0.170)
N	1,698	1,698	1,266	1,690	1,258

Source: Own calculations from the first four waves of NIDS. Attrition-corrected panel weights applied to members of the balanced panel. Bootstrapped standard errors in parentheses.

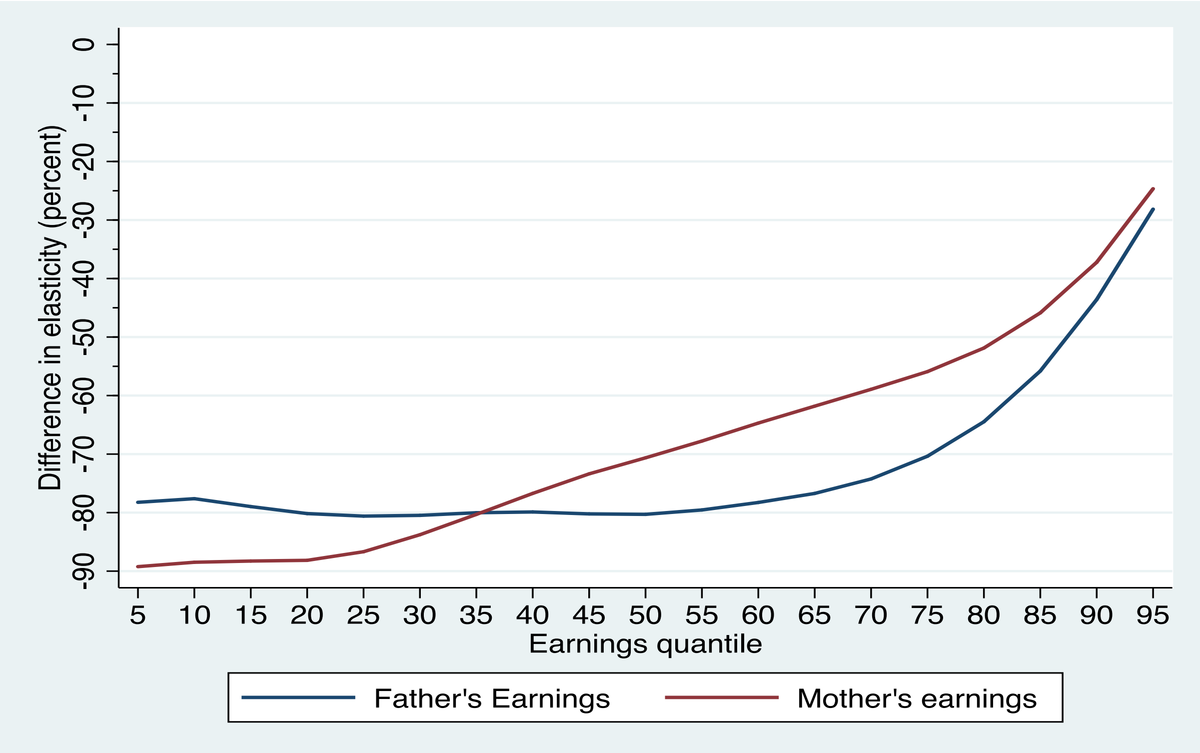
4.D Elasticities for different imputations and different subsamples with double correction

Table 4.D.1: Intergenerational earnings elasticities for different imputation procedures with a double Heckman correction for different subsamples

	Variables used to construct parental earnings				
	1 Education	**2** Education, race	**3** Education, race, occupation	**4** Education, race, province	**5** Education, race, occupation, province
			Father's earnings		
Elasticity	0.641	0.750	0.704	0.742	0.678
	(0.242)	(0.241)	(0.217)	(0.234)	(0.215)
N	1,782	1,782	1,397	1,774	1,389
			Mother's earnings		
Elasticity	0.681	0.767	0.660	0.838	0.718
	(0.250)	(0.261)	(0.215)	(0.218)	(0.220)
N	1,698	1,698	1,266	1,690	1,258

Source: Own calculations from the first four waves of NIDS. Attrition-corrected panel weights applied to members of the balanced panel. Bootstrapped standard errors in parentheses.

Figure 4.D.1: Difference in intergenerational elasticity when controlling for education, over parental earnings quantiles: African balanced panel sample only



Source: Own calculations from the first four waves of NIDS. Attrition-corrected panel weights applied to members of the balanced panel.

5 Conclusion

5.1 Main findings and implications

This aim of this thesis was to investigate three distinct ideas about economic mobility in South Africa in light of the persistently high levels of inequality, poverty and unemployment in the country. Thinking about these as dynamic, rather than static, concepts allowed for novel questions to be asked, and different conclusions to be drawn, about longer-run welfare.

Economic mobility can mean many different things, and so the main contributions of this thesis to the South African and international mobility literature are widespread. The first main contribution is the use of an observed counterfactual state to assess the role of a particular type of measurement error in determining the measurement of mobility. This is combined with the application of a useful method for researchers to assess the validity of their household and individual survey data in a wide variety of settings. The second main contribution is more specific to the South African literature, in that this is the first example of a multi-wave nationally representative study of absolute mobility in the country. A profile of the poor, and of the characteristics associated with poverty, is presented, along with careful econometric modelling to uncover the processes determining poverty entry and poverty persistence. The third main contribution is the extension of the literature on the intergenerational transmission of earnings by going beyond a simple estimation of the intergenerational correlation of earnings. This is done by implementing a modelling strategy that takes account of high unemployment rates in both generations. The result is that controlling for high unemployment makes a significant difference to our understanding of intergenerational mobility, as the shape of the intergenerational earnings correlation over the distribution of wages suggests. This has wider applications in the estimation of the intergenerational transmission of earnings in other societies that have similarly high unemployment levels to South Africa.

Chapter 2 investigated how prevalent data fabrication is in South African household surveys in general, and in NIDS in particular. It then made use of the first two waves of NIDS to ask how the presence of such fabrication would affect the validity of empirical analyses of labour market mobility in the country. The chapter documented how measurement error from fieldworker cheating was identified, and noted that it affected about 7% of the sample. The fabrication was detected while fieldwork was still on-going, and the relevant interviews were re-conducted. There is therefore an observed counterfactual that can be used to measure how problematic such fabrication would have been, had it remained undetected. The chapter compared estimates from the dataset that included the fabricated interviews with corresponding estimates that included the corrected data instead. The results indicated that the fabrication would not have affected our univariate and cross-sectional estimates meaningfully, but would have led us to reach substantially different conclusions when implementing panel estimators. The data quality investigation in this survey had a cost-benefit ratio of at least 24, and was thus easily justifiable.

Chapter 3 analysed the determinants of South Africans moving into and out of poverty over the first four waves of NIDS for the years 2008 to 2014/2015. The first descriptive sections of the chapter focused on the balanced panel of NIDS respondents and found that a relatively high poverty exit rate was accompanied by a substantial proportion of the population being trapped in severe poverty. The roles of demographic versus income changes over time revealed that changing household composition was the largest trigger of poverty entry and exit, and that increasing income from government grants was the main trigger precipitating poverty exit for about one quarter of our sample. The chapter then presented the results from an endogenous switching model that controlled for initial conditions and selective attrition on the full sample of respondents in order to better understand what traps South Africans in poverty. The findings showed that ignoring the correlations between the unobservables affecting initial conditions, sample retention and poverty transitions lead to substantially biased results, and that there was significant genuine state dependence underlying poverty dynamics. This has important policy implications, as preventing people from falling into poverty in the first place is likely to yield greater returns than targeting the individual correlates of poverty directly. Taking a dynamic, rather than static, view of poverty was also useful in examining who would benefit from structural interventions aimed at chronic poverty relief, versus who would benefit from some kind of insurance to cover transitory fluctuations in income.

Chapter 4 asked how the correlation between the earnings of parents and children in South Africa should be calculated in the presence of high unemployment, and what the role of education is in determining this relationship. The chapter used the first four waves of the NIDS and the 1993 PSLSD to investigate the shape of the association between parental and child earnings across the earnings distribution, and found that the correlation was strongest at the ends of the distribution. The estimates were corrected for possible biases that arise from co-resident parent-child pairs, and from selection into labour market participation in South Africa's high-unemployment society. The findings suggest that correcting for selection into employment increased the intergenerational elasticity of earnings by approximately 10 per cent. The role of education in determining the association of intergenerational earnings was uncovered. It was shown that the impact was strongest at the bottom of the earnings distribution, and that education accounted for approximately 40 per cent of the total intergenerational earnings elasticity in the country.

There are a number of potential avenues of research that can build on and extend some of the key findings of this thesis. Combining administrative data with NIDS offers an opportunity to introduce a measure of schooling quality into an analysis of changing returns to education and the determination of intergenerational earnings persistence. Greater access to South African tax data will allow for a calibration of the earnings data in NIDS with an official source, which will allow for a new perspective on earnings inequality, as well as cohort analyses of intergenerational mobility. The addition of future waves of data to NIDS presents opportunities for a richer analysis of earnings and income dynamics. For example, as the time period covered by NIDS increases, researchers will be able to estimate the intergenerational persistence of earnings using a single dataset, rather than having to rely on an auxiliary dataset such as the PSLSD. This will allow not only for cleaner measures of permanent income for parents and children, but also for additional decompositions of the roles of various characteristics in determining earnings and the transmission of economic advantage. Having more waves of data will also allow researchers to analyse household formation and dissolution, and the role of income in driving these dynamics. This is something that the thesis touched upon in Chapter 3, and is rich ground for future research. Another question that could be investigated in order to expand on this thesis is how the large post-apartheid increase in female labour force participation affected poverty persistence and transitions, as well as the intergenerational transmission of earnings from mothers to daughters and mothers to sons. There is also important work to be done on looking at how the differing intervals of time between each wave of NIDS impacts poverty dynamics, and whether the presence of relatively long intervals means that transitory poverty is underestimated. Finally, as more high quality longitudinal data becomes available in developing countries, there is increased scope for cross-national analysis of the different drivers of poverty persistence, poverty transitions and intergenerational earnings.

One of the general lessons from this study is that the labour market is central to understanding the dynamics of poverty and inequality in South Africa. Although expanding social welfare policies has been moderately effective in alleviating money-metric poverty for some South Africans in a static sense, the sustainability of this approach in isolation can be questioned. Any long-run policy proposals to tackle poverty and inequality must be dynamic in nature, and this means focusing on increasing labour market access, and increasing the wages of those who are earning at the bottom end of the distribution. Active labour market policies may serve to increase upward mobility for the same people over time. These need to be implemented in conjunction with reforms to national education policies which can increase upward mobility from generation to generation. The fact that over two decades of democracy in South Africa have not seen a substantial reduction in the inequality of opportunity facing parents and children speaks directly to the low quality of public education in the country. Active labour market policies and a commitment to improving the quality of primary and secondary education must surely be complementary foci if policymakers are to ensure that the benefits of living in a democratic society are to be enjoyed by all South Africans.

References

Abel, M. (2013), Unintended labour supply effects of cash transfer programmes: Evidence from South Africa's old age pension, SALDRU Working Papers 114, Southern Africa Labour and Development Research Unit, University of Cape Town.

Agüero, J., Carter, M. and May, J. (2007), 'Poverty and inequality in the first decade of South Africa's democracy: What can be learnt from panel data from KwaZulu-Natal?', *Journal of African Economies*, 16(5), 782–812.

Aigner, D. J. (1973), 'Regression with a binary independent variable subject to errors of observation', *Journal of Econometrics*, 1(1), 49–59.

Aliber, M. (2003), 'Chronic poverty in South Africa: Incidence, causes and policies', *World Development*, 31(3), 473–490.

Ardington, C., Branson, N., Lam, D. and Leibbrandt, M. (2011), 'Explaining the persistence of racial gaps in schooling in South Africa', *African Population Studies*, 25(2), 509–542.

Ardington, C. and Case, A. (2009), Health: Analysis of the NIDS wave 1 dataset, NIDS Discussion Paper 2, National Income Dynamics Study, University of Cape Town.

Ardington, C., Case, A. and Hosegood, V. (2009), 'Labor supply responses to large social transfers: Longitudinal evidence from South Africa', *American Economic Journal. Applied Economics*, 1(1), 22–48.

Ardington, C. and Gasealahwe, B. (2012), Health: Analysis of the NIDS wave 1 and 2 datasets, SALDRU Working Papers 80, Southern Africa Labour and Development Research Unit, University of Cape Town.

Arulampalam, W., Booth, A. and Taylor, M. (2000), 'Unemployment persistence', *Oxford Economic Papers*, 52(1), 24–50.

Asadullah, M. N. (2011), Intergenerational wealth mobility in rural Bangladesh, IZA Discussion Paper 5914, Institute for the Study of Labor.

Atkinson, A. (2015), *Inequality: What can be done?*, Harvard University Press.

Azomahou, T. and Yitbarek, E. (2015), Poverty persistence and informal risk management: Micro evidence from urban Ethiopia, UNU-MERIT Working Papers 2015-006, UNU-MERIT.

Baigrie, N. and Eyal, K. (2013), 'An evaluation of the determinants and implications of panel attrition in the National Income Dynamics Study (2008-2010)', *South African Journal of Economics*, 82(1), 39–65.

Bane, M. and Ellwood, D. (1986), 'Slipping into and out of poverty: The dynamics of spells', *Journal of Human Resources*, 21(1), 1–23.

Becker, G. and Tomes, N. (1979), 'An equilibrium theory of the distribution of income and intergenerational mobility', *Journal of Political Economy*, 87(6), 1153–1189.

Benford, F. (1938), 'The law of anomalous numbers', *Proceedings of the American Philosophical Society*, 78(4), 551–572.

Bergman, M. and Joye, D. (2005), 'Comparing social stratification schemata: CAMSIS, CSP-CH, Goldthorpe, ISCO-88, Treiman, & Wright', *Cambridge Studies in Social Research*, (10), 1–35.

Bertrand, M., Mullainathan, S. and Miller, D. (2003), 'Public policy and extended families: Evidence from pensions in South Africa', *World Bank Economic Review*, 17(1), 27–50.

Bevis, L. and Barrett, C. (2015), 'Decomposing intergenerational income elasticity: The gender-differentiated contribution of capital transmission in rural Philippines', *World Development*, 74, 233–252.

Bhorat, H., Oosthuizen, M. and van der Westhuizen, C. (2012), 'Estimating a poverty line: An application to free municipal services in South Africa', *Development Southern Africa*, 29(1), 77–96.

Biewen, M. (2009), 'Measuring state dependence in individual poverty histories when there is feedback to employment status and household composition', *Journal of Applied Econometrics*, 24(7), 1095–1116.

Birnbaum, B. (2012), Algorithmic Approaches to Detecting Interviewer Fabrication in Surveys, PhD thesis, University of Washington.

Björklund, A. and Jäntti, M. (1997), 'Intergenerational income mobility in Sweden compared to the United States', *American Economic Review*, 87(4), 1009–1018.

Black, S. and Devereux, P. (2011), Recent developments in intergenerational mobility, *in* O. Ashenfelter and D. Card, eds, 'Handbook of Labor Economics', Vol. 4B, Elsevier, chapter 16, pp. 1487–1541.

Blanden, J. (2009), How much can we learn from international comparisons of intergenerational mobility?, CEE Discussion Paper 111, Centre for the Economics of Education.

Blanden, J. (2015), 'Intergenerational income persistence', *IZA World of Labor 2015*, 176.

Blanden, J., Gregg, P. and Macmillan, L. (2007), 'Accounting for intergenerational income persistence: Noncognitive skills, ability and education', *The Economic Journal*, 117(519), C43–C60.

Blanden, J. and Macmillan, L. (2014), Education and intergenerational mobility: Help or hindrance?, DoQSS Working Paper 14-01, Department of Quantitative Social Science, Institute of Education, University of London.

Boskin, M. and Nold, F. (1975), 'A Markov model of turnover in aid to families with dependent children', *The Journal of Human Resources*, 10(4), 467–481.

Bound, J., Brown, C. and Mathiowetz, N. (2001), Measurement error in survey data, *in* J. Heckman and E. Leamer, eds, 'Handbook of Econometrics', Vol. 5 of *Handbook of Econometrics*, Elsevier, chapter 59, pp. 3705–3843.

Bowles, S. and Gintis, H. (2002), 'The inheritance of inequality', *Journal of Economic Perspectives*, 16(3), 3–30.

Branson, N., Hofmeyr, C. and Lam, D. (2014), 'Progress through school and the determinants of school dropout in South Africa', *Development Southern Africa*, 31(1), 106–126.

Branson, N. and Leibbrandt, M. (2013), Education quality and labour market outcomes in South Africa, OECD Working Papers 1021, OECD Economics Department.

Bratberg, E., Nilsen, Ø. and Vaage, K. (2007), 'Trends in intergenerational mobility across offspring's earnings distribution in Norway', *Industrial Relations*, 46(1), 112–129.

Bredl, S., Winker, P. and Kötschau, K. (2008), A statistical approach to detect cheating interviewers, Technical Report 39, Diskussionsbeiträge: Zentrum für internationale Entwicklungs-und Umweltforschung.

Brunori, P., Ferreira, F. and Peragine, V. (2013), Inequality of opportunity, income inequality, and economic mobility: Some international comparisons, *in* E. Paus, ed., 'Getting development right: Structural transformation, inclusion, and sustainability in the post-crisis era', Palgrave Macmillan, chapter 5, pp. 85–115.

Buchinsky, M. (1998), 'Recent advances in quantile regression models: A practical guideline for empirical research', *The Journal of Human Resources*, 33(1), 88–126.

Buddelmeyer, H. and Verick, S. (2008), 'Understanding the drivers of poverty dynamics in Australian households', *Economic Record*, 84(266), 310–321.

Budlender, J., Leibbrandt, M. and Woolard, I. (2015), South African poverty lines: A review and two new money-metric thresholds, SALDRU Working Papers 151, Southern Africa Labour and Development Research Unit, University of Cape Town.

Burger, R., Klasen, S. and Zoch, A. (2016), Estimating income mobility when income is measured with error: The case of South Africa, ERSA Working Papers 607, Economic Research Southern Africa.

Cappellari, L. and Jenkins, S. P. (2004), 'Modelling low income transitions', *Journal of Applied Econometrics*, 19(5), 593–610.

Cappellari, L. and Jenkins, S. P. (2006), 'Calculation of multivariate normal probabilities by simulation, with applications to maximum likelihood estimation', *The Stata Journal*, 6(2), 156–189.

Carslaw, C. A. (1988), 'Anomalies in income numbers: Evidence of goal oriented behavior', *Accounting Review*, 63(2), 321–327.

Carter, M. and May, J. (2001), 'One kind of freedom: Poverty dynamics in post-apartheid South Africa', *World Development*, 29(12), 1987–2006.

Casale, D. and Posel, D. (2002), 'The continued feminisation of the labour force in South Africa', *South African Journal of Economics*, 70(1), 156–184.

Cervini-Plá, M. (2013), 'Exploring the sources of earnings transmission in Spain', *Review of Public Economics*, 204(1), 45–66.

Cervini-Plá, M. (2015), 'Intergenerational earnings and income mobility in Spain', *Review of Income and Wealth*, 61(4), 812–828.

Checchi, D. (1997), 'Education and intergenerational mobility in occupations: A comparative study', *American Journal of Economics and Sociology*, 56(3), 331–351.

Chinhema, M., Brophy, T., Brown, M., Leibbrandt, M., Mlatsheni, C. and Woolard, I. (2016), National Income Dynamics Study Panel User Manual, Technical report, Southern Africa Labour and Development Research Unit.

Cho, M., Eltinge, J. and Swanson, D. (2003), Inferential methods to identify possible interviewer fraud using leading digit preference patterns and design effect matrices, *in* 'Proceedings of the American Statistical Association (Survey Research Methods Section)', pp. 936–941.

Chusseau, N., Hellier, J. and Ben-Halima, B. (2013), Education, intergenerational mobility and inequality, *in* J. Hellier and N. Chusseau, eds, 'Growing income inequalities', Palgrave Macmillan, chapter 8, pp. 227–273.

Cichello, P., Fields, G. and Leibbrandt, M. (2005), 'Earnings and employment dynamics for Africans in post-apartheid South Africa: A panel study of KwaZulu-Natal', *Journal of African Economies*, 14(2), 143–190.

Clarke, R. and Eyal, K. (2014), 'Microeconomic determinants of spatial mobility in post-apartheid South Africa: Longitudinal evidence from the National Income Dynamics Study', *Development Southern Africa*, 31(1), 168–194.

Conlisk, J. (1971), 'A bit of evidence on the income-education-ability interrelation', *Journal of Human Resources*, 6(3), 358–362.

Conlisk, J. (1974), 'Can equalization of opportunity reduce social mobility', *American Economic Review*, 64(1), 80–90.

Corak, M. (2013), Income inequality, equality of opportunity, and intergenerational mobility, IZA Discussion Paper 7520, Institute for the Study of Labor.

de Villiers, L., Brown, M., Woolard, I. and Leibbrandt, M. (2013), National Income Dynamics Study Wave 3 user manual, Technical report, Southern Africa Labour and Development Research Unit.

Deaton, A. (1997), *The analysis of household surveys: A microeconometric approach to development policy*, World Bank Publications.

Devey, R., Skinner, C. and Valodia, I. (2006), Definitions, data and the informal economy in South Africa: a critical analysis, *in* V. Padayachee, ed., 'The Development Decade?: Economic and Social Change in South Africa, 1994-2004', HSRC Press, chapter 15, pp. 302–323.

Devicienti, F. and Poggi, A. (2011), ‘Poverty and social exclusion: Two sides of the same coin or dynamically interrelated processes?’, *Applied Economics*, 43(25), 3549–3571.

Doan, Q. and Nguyen, N. (2016), Intergenerational income mobility in Vietnam, MPRA Paper 70603, University Library of Munich.

DSD, SASSA and UNICEF (2012), The South African child support grant impact assessment: Evidence from a survey of children, adolescents and their households, Technical report, UNICEF South Africa.

Durtschi, C., Hillison, W. and Pacini, C. (2004), ‘The effective use of Benford’s law to assist in detecting fraud in accounting data’, *Journal of Forensic Accounting*, 5(1), 17–34.

Eide, E. and Showalter, M. (1999), ‘Factors affecting the transmission of earnings across generations: A quantile regression approach’, *The Journal of Human Resources*, 34(2), 253–267.

Erikson, R. and Goldthorpe, J. (1992), *The constant flux: A study of class mobility in industrial societies*, Oxford University Press.

Ermisch, J., Francesconi, M. and Siedler, T. (2006), ‘Intergenerational mobility and marital sorting’, *The Economic Journal*, 116(513), 659–679.

Essers, D. (2017), ‘South African labour market transitions since the global financial and economic crisis: Evidence from two longitudinal datasets’, *Journal of African Economies*, 26(2), 192–222.

Eyal, K. and Burns, J. (2016), Up or down? Intergenerational mental health transmission and cash transfers in South Africa, SALDRU Working Papers 165, Southern Africa Labour and Development Research Unit, University of Cape Town.

Faye, O., Islam, N. and Zulu, E. (2011), Poverty dynamics in Nairobi’s slums: Testing for true state dependence and heterogeneity effects, CEPS/INSTEAD Working Papers 2011-56, CEPS/INSTEAD.

Ferreira, S. and Veloso, F. (2006), ‘Intergenerational mobility of wages in Brazil’, *Brazilian Review of Econometrics*, 26(2), 181–211.

Fields, G. (2010), But that's not what economic mobility is!, Cornell University, ILR School Working Papers 154, Cornell University, ILR School.

Finn, A. (2015), A national minimum wage in the context of the South African labour market, SALDRU Working Papers 153, Southern Africa Labour and Development Research Unit, University of Cape Town.

Finn, A. and Leibbrandt, M. (2013), The dynamics of poverty in the first three waves of NIDS, SALDRU Working Papers 119, Southern Africa Labour and Development Research Unit, University of Cape Town.

Finn, A. and Leibbrandt, M. (2016), The dynamics of poverty in the first four waves of NIDS, SALDRU Working Papers 174, Southern Africa Labour and Development Research Unit, University of Cape Town.

Finn, A., Leibbrandt, M. and Levinsohn, J. (2013), 'Income mobility in a high-inequality society: Evidence from the National Income Dynamics Study', *Development Southern Africa*, 4(6).

Finn, A., Leibbrandt, M. and Oosthuizen, M. (2014), Poverty, inequality and prices in post-apartheid South Africa, WIDER Working Papers 2014/127, World Institute for Development Economics Research.

Finn, A., Leibbrandt, M. and Ranchhod, V. (2014), Post-apartheid poverty and inequality trends, *in* R. Kanbur, ed., 'The Oxford Companion to the Economics of South Africa', Oxford University Press.

Friedman, M. (1962), *Capitalism and freedom*, University of Chicago Press.

Gong, H., Leigh, A. and Meng, X. (2012), 'Intergenerational income mobility in urban China', *Review of Income and Wealth*, 58(3), 481–503.

Gouriéroux, C. and Monfort, A. (1996), *Simulation based econometric methods*, Oxford University Press.

Grawe, N. (2004), Intergenerational mobility for whom? The experience of high- and low-earning sons in international perspective, *in* M. Corak, ed., 'Generational Income Mobility in North America and Europe', Cambridge University Press, chapter 4, pp. 58–89.

Griliches, Z. and Hausman, J. A. (1986), 'Errors in variables in panel data', *Journal of Econometrics*, 31(1), 93–118.

Gronau, R. (1974), 'Wage comparisons - A selectivity bias', *Journal of Political Economy*, 82(6), 1119–1144.

Haile, G. (2016), Intergenerational mobility in income and economic status in Ethiopia, IZA Discussion Paper 10047, Institute for the Study of Labor.

Hausman, J. (2001), 'Mismeasured variables in econometric analysis: Problems from the right and problems from the left', *Journal of Economic Perspectives*, 15(4), 57–67.

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

Heckman, J. (1981*a*), The incidental parameters problem and the problem of initial conditions in estimating a discrete time - discrete data stochastic process, *in* C. Manski and D. McFadden, eds, 'Structural Analysis of Discrete Data with Econometric Applications', MIT Press, chapter 4, pp. 179–195.

Heckman, J. (1981*b*), Statistical models for discrete panel data, *in* C. Manski and D. McFadden, eds, 'Structural Analysis of Discrete Data with Econometric Applications', MIT Press, chapter 3, pp. 114–178.

Hertz, T. (2001), Education, inequality and economic mobility in South Africa, PhD thesis, University of Massachusetts.

Hill, T. P. (1995), 'A statistical derivation of the significant-digit law', *Statistical Science*, 10(4), 354–363.

Hnatovska, V., Lahiri, A. and Paul, S. (2013), 'Breaking the caste barrier: Intergenerational mobility in India', *Journal of Human Resources*, 48(2), 435–473.

Hulme, D. and Shepherd, A. (2003), 'Conceptualizing chronic poverty', *World Development*, 31(3), 403–423.

International Labour Office (2011), Unemployment, underemployment and inactivity indicators, *in* 'Key indicators of the labour market', ILO, chapter 4.

Jalan, J. and Ravallion, M. (2000), 'Is transient poverty different? Evidence for rural china', *Journal of Development Studies*, 36(6), 82–99.

Jäntti, M., Bratsberg, B., Røed, K., Raaum, O., Naylor, R., Österbacka, E., Björklund, A. and Eriksson, T. (2006), American exceptionalism in a new light: A comparison of intergenerational earnings mobility in the Nordic countries, the United Kingdom and the United States, IZA Discussion Paper 1938, Institute for the Study of Labor.

Jäntti, M. and Jenkins, S. P. (2013), Income mobility, IZA Discussion Paper 7730, Institute for the Study of Labor.

Jarvis, S. and Jenkins, S. P. (1997), 'Low income dynamics in 1990s Britain', *Fiscal Studies*, 18(2), 123–142.

Jenkins, S. P. (2011), *Changing fortunes: Income mobility and poverty dynamics in Britain*, Oxford University Press.

Jenkins, S. P. and Rigg, J. A. (2001), The dynamics of poverty in Britain, Research Report 157, Department for Work and Pensions.

Judge, G. and Schechter, L. (2009), 'Detecting problems in survey data using Benford's Law', *Journal of Human Resources*, 44(1), 1–24.

Kerr, A. and Wittenberg, M. (2016), A guide to version 3.1 of the post-apartheid labour market series (PALMS), PALMS User Guide 3.1, DataFirst, University of Cape Town.

Keswell, M., Girdwood, S. and Leibbrandt, M. (2013), 'Educational inheritance and the distribution of occupations: Evidence from South Africa', *Review of Income and Wealth*, 59, S111–S137.

Keswell, M. and Poswell, L. (2004), ‘Returns to education in South Africa: A retrospective sensitivity analysis of the available evidence’, *South African Journal of Economics*, 72(4), 834–860.

Klevmarken, N. A. (1982), ‘Missing variables and two-stage least-squares estimation from more than one dataset’, *1981 Proceedings of the American Statistical Association, Business and Economic Statistics Section*, pp. 156–161.

Krueger, A. B. and Summers, L. H. (1988), ‘Efficiency wages and the inter-industry wage structure’, *Econometrica*, 56(2), 259–93.

Lam, D., Ardington, C., Branson, N. and Leibbrandt, M. (2013), Credit constraints and the racial gap in post-secondary education in South Africa, NBER Working Paper 19607, National Bureau of Economic Research.

Lam, D., Ardington, C., Branson, N., Maughan-Brown, B., Menendez, A., Seekings, J. and Sparks, M. (2012), The Cape Area Panel Study: Overview and technical documentation waves 1-2-3-4-5(2002-2009), Technical report, The University of Cape Town.

Lam, D., Finn, A. and Leibbrandt, M. (2015), Schooling inequality, returns to schooling, and earnings inequality: Evidence from Brazil and South Africa, WIDER Working Paper 2015/050, United Nations University World Institute for Development Economics Research.

Lee, S., Linton, O. and Whang, Y.-J. (2009), ‘Testing for stochastic monotonicity’, *Econometrica*, 77(2), 585–602.

Lefranc, A. and Trannoy, A. (2005), ‘Intergenerational earnings mobility in France: Is France more mobile than the US?’, *Annales d’Economie et Statistique*, 78, 57–77.

Leibbrandt, M., Finn, A. and Oosthuizen, M. (2016), Poverty, inequality and prices in post-apartheid South Africa, *in* C. Arndt, A. McKay and F. Tarp, eds, ‘Growth and Poverty in Sub-Saharan Africa’, Oxford: Oxford University Press, chapter 17, pp. 393–417.

Leibbrandt, M., Finn, A. and Woolard, I. (2012), ‘Describing and decomposing post-apartheid income inequality in South Africa’, *Development Southern Africa*, 29(1), 19–34.

Leibbrandt, M. and Levinsohn, J. (2016), Fifteen years on: Household incomes in South Africa, *in* S. Edwards, S. Johnson and D. Weil, eds, 'African Successes, Volume I: Government and Institutions', National Bureau of Economic Research, chapter 9, pp. 333–355.

Leibbrandt, M., Woolard, I., Finn, A. and Argent, J. (2010), Trends in South African income distribution and poverty since the fall of apartheid, OECD Social, Employment and Migration Working Papers 101, OECD Publishing.

Li, J., Brick, J., Tran, B. and Singer, P. (2011), 'Using statistical models for sample design of a reinterview program', *Journal of Official Statistics*, 27(3), 433–450.

Louw, M., van der Berg, S. and Yu, D. (2007), 'Convergence of a kind: Educational attainment and intergenerational social mobility in South Africa', *South African Journal of Economics*, 75(3), 548–571.

May, J., Agüero, J., Carter, M. and Timaeus, I. (2007), 'The KwaZulu-Natal Income Dynamics Study (KIDS) 3rd wave: Methods, first findings and an agenda for future research', *Development Southern Africa*, 24(5), 629–648.

Murphy, J., Baxter, R., Eyerman, J., Cunningham, D. and Kennet, J. (2004), A system for detecting interviewer falsification, *in* 'American Association for Public Opinion Research, 59th Annual Conference', pp. 4968–4975.

National Treasury (2016), A national minimum wage for South Africa: Recommendations on policy and implementation, National minimum wage panel report to the deputy president, National Treasury.

Nigrini, M. (1999), 'I've got your number', *Journal of Accountancy*, 187(5), 79–83.

Nilsen, Ø., Vaage, K., Aakvik, A. and Jacobsen, K. (2012), 'Intergenerational earnings mobility revisited: Estimates based on lifetime earnings', *Scandinavian Journal of Economics*, 114(1), 1–23.

Nybom, M. and Stuhler, J. (2016), 'Heterogeneous income profiles and lifecycle bias in intergenerational mobility estimation', *Journal of Human Resources*, 51(1), 239–268.

Österbacka, E. (2001), 'Family background and economic status in Finland', *Scandinavian Journal of Economics*, 103(3), 467–484.

Özler, B. (2007), 'Not separate, not equal: Poverty and inequality in post-apartheid South Africa', *Economic Development and Cultural Change*, 55, 487–529.

Palomino, J., Marrero, G. and Rodriguez, J. (2014), One size doesn't fit all: A quantile analysis of intergenerational income mobility in the US (1980-2010), ECINEQ Working Paper 349, Society for the Study of Economic Inequality.

Pellicer, M., Piraino, P. and Wegner, E. (2014), Information, mobilization, and demand for redistribution: A survey experiment in South Africa, SALDRU Working Papers 139, Southern Africa Labour and Development Research Unit, University of Cape Town.

Piraino, P. (2015), 'Intergenerational earnings and inequality of opportunity in South Africa', *World Development*, 67, 396–405.

Porras, J. and English, N. (2004), Data-driven approaches to identifying interviewer data falsification: The case of health surveys, *in* 'Proceedings of the American Statistical Association (Survey Research Methods Section)', pp. 4223–4228.

Posel, D. (2010), 'Households and labour migration in post-apartheid South Africa', *Studies in Economics and Econometrics*, 34(3), 129–141.

Posel, D. and Casale, D. (2003), 'What has been happening to internal labour migration in South Africa, 1993-1999?', *South African Journal of Economics*, 71(3), 455–479.

Posel, D., Fairburn, J. and Lund, F. (2006), 'Labour migration and households: A reconsideration of the effects of the social pension on labour supply in South Africa', *Economic Modelling*, 23(5), 836–953.

Posel, D. and Rogan, M. (2016), 'Measured as poor versus feeling poor: Comparing money-metric and subjective poverty rates in South Africa', *Journal of Human Development and Capabilities*, 17(1), 55–73.

Ranchhod, V. (2009), Household responses to adverse income shocks: Pensioner out-migration and mortality in South Africa, SALDRU Working Papers 35, Southern Africa Labour and Development Research Unit, University of Cape Town.

Ranchhod, V. and Finn, A. (2015), Estimating the effects of South Africa's youth employment tax incentive - An update, SALDRU Working Papers 152, Southern Africa Labour and Development Research Unit, University of Cape Town.

Ranchhod, V. and Finn, A. (2016), 'Estimating the short run effects of South Africa's employment tax incentive on youth employment probabilities using a difference-in-differences approach', *South African Journal of Economics*, 84(2), 199–216.

Ransom, R. and Sutch, R. (1977), *One kind of freedom: The economic consequences of emancipation*, Cambridge University Press.

Roberts, B. (2001), 'Chronic and transitory poverty in post-apartheid South Africa: Evidence from KwaZulu-Natal', *Journal of Poverty*, 5(4), 1–28.

SALDRU (2016*a*), 'National Income Dynamics Study 2008, Wave 1 [dataset]. Version 5.3', Cape Town: Southern Africa Labour and Development Research Unit [producer], Cape Town: DataFirst [distributor].

SALDRU (2016*b*), 'National Income Dynamics Study 2010-2011, Wave 2 [dataset]. Version 2.3', Cape Town: Southern Africa Labour and Development Research Unit [producer], Cape Town: DataFirst [distributor].

SALDRU (2016*c*), 'National Income Dynamics Study 2012, Wave 3 [dataset]. Version 1.3', Cape Town: Southern Africa Labour and Development Research Unit [producer], Cape Town: DataFirst [distributor].

SALDRU (2016*d*), 'National Income Dynamics Study 2014-2015, Wave 4 [dataset]. Version 1.0', Cape Town: Southern Africa Labour and Development Research Unit [producer], Cape Town: DataFirst [distributor].

Schäfer, C., Schräpler, J.-P., Müller, K.-R. and Wagner, G. G. (2004), Automatic identification of faked and fraudulent interviews in surveys by two different methods, DIW Discussion Paper 441, DIW Berlin, German Institute for Economic Research.

Schnell, R. (1991), 'Der Einfluß gefälschter Interviews auf Survey Ergebnisse', *Zetischrift für Soziologie*, 20, 25–35.

Schnitzlein, D. (2016), 'A new look at intergenerational mobility in Germany compared to the US', *Review of Income and Wealth*, 62(4), 650–667.

Schräpler, J.-P. (2011), 'Benford's law as an instrument for fraud detection in surveys using the data of the socio-economic panel (SOEP)', *Journal of Economics and Statistics*, 231(5-6), 685–718.

Schräpler, J.-P. and Wagner, G. (2005), 'Characteristics and impact of faked interviews in surveys - An analysis of genuine fakes in the raw data of SOEP', *Allgemeines Statistisches Archiv*, 89(1), 7–20.

Schreiner, I., Pennie, K. and Newbrough, J. (1988), Interviewer falsification in census bureau surveys, *in* 'Proceedings of the American Statistical Association (Survey Research Methods Section)', pp. 491–496.

Scott, P. D. and Fasli, M. (2001), Benford's law: An empirical investigation and a novel explanation, CSM Technical Report 349, Department of Computer Science, University of Essex.

Sienaert, A. (2008), The labour supply effects of the South African old age pension: Theory, evidence and implications, SALDRU Working Papers 20, Southern Africa Labour and Development Research Unit, University of Cape Town.

Solon, G. (1992), 'Intergenerational income mobility in the United States', *American Economic Review*, 82(3), 393–408.

South African Reserve Bank (2017), 'Quarterly Bulletin Online Statistical Query', Online database. Available at www.reservebank.co.za. Accessed August 2017.

Standish-White, J. and Finn, A. (2015), Unconditional cash transfers and children's educational outcomes: Evidence from the old age pension programme in South Africa, SALDRU Working Papers 147, Southern Africa Labour and Development Research Unit, University of Cape Town.

Statistics South Africa (2000), Time Use Survey: Fieldworker's manual.

Statistics South Africa (2015), Methodological report on rebasing of national poverty lines and development of pilot provincial poverty lines, Technical report 03-10-11, Statistics South Africa.

Stewart, M. and Swaffield, J. (1999), 'Low pay dynamics and transition probabilities', *Economica*, 66(261), 23–42.

Swanson, D., Cho, M. and Eltinge, J. (2003), Detecting possibly fraudulent or error-prone survey data using Benford's law, *in* 'Proceedings of the American Statistical Association (Survey Research Methods Section)', pp. 4172–4177.

Thomas, J. K. (1989), 'Unusual patterns in reported earnings', *The Accounting Review*, 64(4), 773–787.

van der Berg, S., Burger, C., Burger, R., de Vos, M., du Rand, G., Gustafsson, M., Moses, E., Shepherd, D., Spaull, N., Taylor, S., van Broekhuizen, H. and von Fintel, D. (2011), Low quality education as a poverty trap, Stellenbosch Economic Working Papers 25/11, University of Stellenbosch.

Vella, F. (1998), 'Estimating models with sample selection bias: A survey', *The Journal of Human Resources*, 33(1), 127–169.

von Fintel, M., Zoch, A. and van der Berg, S. (2016), 'The dynamics of child poverty in South Africa between 2008 and 2012', *Child Indicators Research*, pp. 1–25.

Wittenberg, M. (2009), Weighing the value of asset proxies: The case of the body mass index in South Africa, SALDRU Working Papers 39, Southern Africa Labour and Development Research Unit, University of Cape Town.

Wittenberg, M. (2013), ‘The weight of success: The body mass index and economic well-being in Southern Africa’, *Review of Income and Wealth*, 59, S62–S83.

Woolard, I. and Klasen, S. (2005), ‘Determinants of income mobility and household poverty dynamics in South Africa’, *The Journal of Development Studies*, 41(5), 865–897.

Woolard, I. and Leibbrandt, M. (2006), Towards a poverty line for South Africa: Background note, Technical report, Southern Africa Labour and Development Research Unit.

Wooldridge, J. M. (2002), *Econometric analysis of cross section and panel data*, The MIT press.

Yu, D. (2010), Poverty and inequality trends in South Africa using different survey data, Stellenbosch Economic Working Papers 04/10, Stellenbosch University.