Incomes in South Africa since the fall of Apartheid

by

Murray Leibbrandt
University of Cape Town
James Levinsohn
University of Michigan
National Bureau of Economic Research
Justin McCrary
University of Michigan

Current version: May 19, 2005

Abstract. This paper examines changes in individual real incomes in South Africa between 1995 and 2000. We document substantial declines—on the order of 40%—in real incomes for both men and women. The brunt of the income decline appears to have been shouldered by the young and the non-White. We extend nonparametric methodologies to examine the role of changes in endowments, returns to these endowments and selection into and out of positive incomes as possible explanations for this income change. We argue that changes in respondent attributes are insufficient to explain this decline. For most groups, a (conservative) correction for selection into income recipiency explains some, but not all, of the income decline. For other groups, selection is a potential explanation for the income decline. Perhaps the most persuasive explanation of the evidence is substantial economic restructuring of the South African economy in which wages are not bid up to keep pace with price changes due to a differentially slack labor market.

Address. McCrary and Levinsohn: Ford School of Public Policy, University of Michigan, Ann Arbor, MI 48109; Leibbrandt: University of Cape Town Private Bag, Rondebosch, 7700 South Africa

Incomes in South Africa since the fall of Apartheid

Murray Leibbrandt
University of Cape Town
James Levinsohn
University of Michigan
National Bureau of Economic Research
Justin McCrary
University of Michigan

1. Introduction and Overview

This paper is about trying to understand how and why the distribution of individual incomes in South Africa has changed since the fall of apartheid. The topic is timely as South Africa recently celebrated the ten year anniversary of democracy in that country. While South Africans clearly have much more political freedom than they did under apartheid, the improvements in economic well-being are less apparent. We employ detailed income and expenditure data from 1995 and from 2000 to examine the latter.

In the normal course of events, one would not expect to see significant changes in the overall distribution of income over only a five year span.[1] The five years following the advent of democracy, though, are not representative of the normal course of events. Even an abbreviated list of changes during this period would include South Africa's re-engagement with the international economy, continued attempts to integrate the apartheid-defined homelands with the new national economy, the possible emigration of highly skilled Whites, and an HIV infection rate that was probably among the highest in the world. All of the above changes have labor market implications. This list, though, is exclusive of government programs focused specifically on the labor market. There were in fact several such programs.

The ruling party in South Africa during this period (and still today) was the African National Congress (ANC). Elected on April 27, 1994, the ANC proceeded to introduce several labor market policies with a focus on redistribution. (See Maziya (2001) for details). In 1995 the new Labour

We are grateful to John DiNardo for countless hours of discussion—some of it even on-topic. Thanks to Jere Behrman, Chico Ferreira, David Lam, Doug Miller, and Berk Ozler for helpful comments. Matthew Welch of DataFirst at the University of Cape Town and Justin Thomas have provided almost six years of invaluable assistance with the data sets. Vimal Ranchhod provided exemplary research assistance. Leibbrandt and Levinsohn gratefully acknowledge research support from the National Institutes of Health.

[1] Bourguignon, Ferreira, and Lustig (2001), in their review of income distribution dynamics, recommend at least a ten year interval.

Relations Act was passed. This law defended the right of workers to unionize and provided a framework for employer-employee dispute resolution. In 1997, the Basic Conditions of Employment Act was passed. This law provided the right to annual leave and imposed rules and procedures to prevent unfair dismissal. In 1998, the Skills Development Act and the Employment Equity Act were passed. The former provided an institutional framework for improving workforce skills while the latter was the centerpiece of the ANC's affirmative action strategy. The Employment Equity Act required all firms of a certain size and turnover to comply with various affirmative action requirements. These included an increase in the proportion of non-Whites and females at all levels. Firms also had to submit a strategic plan, with numerical targets, for increasing the diversity of their workforce. Finally, in 1999, the Skills Development Levies Act was passed and this effectively introduced a payroll tax on employees to help fund the "Skills Development Fund."

Clearly, the labor market was in transition. Pro-labor groups tended to hail these changes as a great and long-past-due success while business often saw the changes as making a bad situation worse. Business groups wondered how taxing labor and increasing union strength would help increase employment. The ANC's labor market policies were concurrent with a macroeconomic policy (coined GEAR– Growth, Employment, and Redistribution) that placed a strong emphasis on macroeconomic stability and export-led growth. The strategy (mostly realized) was to use fiscal conservatism, reduced state debt, and increased openness to the global market to achieve (mostly unrealized) growth.

There were also changes in non-labor market incomes. Prior to around 1992, government old-age pensions were effectively race-based (see Case and Deaton (1998)). By 1995, the first year of our data, government old-age pensions were available to all on a more or less race-blind basis. Hence, changes in the real income provided by these pensions impacted large segments of the population. Accordingly, we investigate the possible role of changes in non-labor market incomes in determining the distribution of overall incomes.

We begin our analysis by simply documenting the changes in individual incomes in South Africa from 1995 to 2000. We are, we believe, the first to do so. Hoogeveen and Ozler (2004) examine changes in *household* income and expenditure, while Lam and Leibbrandt (2004) examine changes in household income and individual *labor market* earnings with a focus on documenting changes in inequality.

We find that the entire distribution of individual log real incomes shifted substantially to the left from 1995 to 2000. This is true for both men and women, and it is true for both labor and (to a lesser extent) pension income. Simply put, individual incomes seem to have fallen in post-apartheid

South Africa. The more interesting question, and one at which we attempt a first pass, is "Why?"[2]

Our analysis does not, in the end, deliver a definitive answer to this provocative question. One explanation that is consistent with the data is a mix of skill-biased technical change, increases in the supply of labor that have outstripped increases in the demand for labor, and growing disparities in the quality of education. There may be other explanations that are also consistent with the data. We show, though, that several potential explanations simply are not consistent with the data. A contribution of this paper is that it presents basic facts with which any explanation for the decline must grapple.

Individual incomes, conditional on being positive, are a function of the individual's endowments (which could include household-level attributes) and returns on those endowments. There are, then, three possible explanations for the change in the distribution of log income. First, endowments might have changed. This could occur either because those with positive incomes are different in the two time periods (this is often termed "selection on observables"), or because endowments are actually changing over time (e.g., human capital). Second, returns to those endowments might have changed. Third and perhaps less obviously, selection into the class of individuals reporting positive income might have changed in ways unrelated to endowments ("selection on unobservables"). We refer to these respectively as the endowments explanation, the returns explanation, and the selection explanation. They are of course not mutually exclusive and, in principle, all could be important.

Our approach to investigating these explanations is, for the most part, nonparametric. To investigate the endowments explanation, we follow the methodology developed by DiNardo, Fortin, and Lemieux (1996), hereinafter DFL. This methodology allows one to investigate the role of changes in the entire vector of endowments as well as the impact of individual components comprising the vector of endowments. We find that while endowments in South Africa experienced some minor changes during our study period, the changes are not sufficient to explain the decline in income.

To investigate the returns explanation, we show how the basic idea behind DFL can be applied to investigate changes in returns (as opposed to endowments). This allows us to nonparametrically investigate the extent to which changes in the entire vector of returns can explain the shift in the distributions of income.[3] However, this technique does not allow one to answer how a shift in the return to only one endowment affected the income distribution, holding constant the return to

[2] We are sensitive to the Holland (1986) critique of such questions. We stress that the key aim of the present paper is description.

[3] We also develop a second extension of the DFL procedure which we argue provides useful context for interpretation of the endowments counterfactual, as proposed in DFL. See section 4, below.

other endowments. The problem does not lie with the methodology, but rather with the model-specific nature of the concept of a return. Consequently, we also consider changes in individual returns in the context of a group-specific linear regression model. Our primary conclusion from the non-parametric approach is that changes in returns can in fact explain a substantial portion of the shift in the distribution of incomes. The more parametric investigation further suggests that young South Africans and Black South Africans have borne the brunt of the recent declines in real income. However, the return to a year of education, as conventionally measured, experienced very little change from 1995 to 2000.

To investigate the selection explanation, we first examine group-specific income changes and compare these magnitudes to patterns in labor market entry and exit. These results indicate that if selection is responsible for some of the estimated income decline, then it is selection of a relatively complex variety. For example, women systematically entered, and men systematically exited, the labor market between 1995 and 2000, and yet both genders experienced real income declines of approximately 40%. We also implement the bounding procedure outlined in Lee (2004),[4] but, to our knowledge, never before actually implemented. For some groups we consider, the results are consistent with a role for selection in explaining the erosion of real incomes; however, for most population groups, even this conservative approach cannot fully explain the recent income declines.

While the focus of this paper is on what happened to incomes in South Africa, the general approach we take may be of interest to a broader literature. Investigating two of the possible explanations (the returns explanation and the selection explanation) required developing and/or implementing new tools to nonparametrically evaluate why distributions of an outcome variable (in this case log income) have shifted. These methodological innovations are applicable to a wide range of problems in economics as heterogeneity in micro-level outcomes is the topic of several fields of economics (e.g., incomes for individuals, productivity for plants, expenditure for households, and exports for firms).

In broad strokes, our analysis supports the view that an individual sampled at random in South Africa was (economically) better off in 1995 than in 2000. We interpret that contrast to be the relevant one for evaluating social welfare, abstracting, of course, from improvements in political freedoms and "psychic income" generally. The remainder of the paper is organized as follows. Section 2 introduces the data. Section 3 sets the stage by describing the patterns in the data that motivate both the questions asked and the methodologies employed. Section 4 describes

[4] Lee's approach is based on Heckman (1976) and Gronau (1974) and leans heavily on the one-sided selection model.

our nonparametric approaches to investigating the endowments, the returns, and the selection explanations. Methodology as well as results are presented in each of three sub-sections—one for each explanation. Section 5 discusses the results, and Section 6 concludes.

2. The Data

Before presenting even descriptive analyses that motivate our nonparametric approach, it is necessary to introduce and describe the data.

We employ five data sets. For 1995, we use the October Household Survey (OHS) and the Income and Expenditure Survey (IES). For 2000, we use the September Labour Force Survey and the 2000 Income and Expenditure Survey. In order to make real comparisons, we employ a data set of price deflators provided by Statistics South Africa.

The 1995 OHS is a general purpose, nationally representative household survey conducted by Statistics South Africa. The OHS is an annual survey that ran from 1994 to 1999. Although the survey instrument is very consistent from year to year, the OHS is not a panel. In 1995, the OHS obtained responses for almost 131,000 individuals comprising about 30,000 households. The OHS is excellent for some purposes and less ideal for others. It has excellent socio-demographic information but the income data in the OHS are notoriously poor. Fortunately for our purposes, the sample for the 1995 OHS was also the sample for the 1995 IES.

The 1995 IES was collected (also by Statistics South Africa) in December, 1995. Of the 29,700 households in the OHS, the IES included 29,582. These households were comprised of almost 129,000 individuals. Not surprisingly, the income information in the IES is very detailed and quite complete.[5] Although the same individuals were in principle included in the OHS and IES, there is entry and exit from some households over the span of October to December. We only include individuals who were present in both the IES and OHS.[6] The careful matched merge of the IES and OHS results in a sample size of over 113,000 individuals from almost 28,000 households.

In 2000, the OHS was replaced by the Labour Force Survey (LFS). The LFS is conducted twice a year and we use the September survey since it is temporally closest to the IES. Perhaps because the LFS is the successor to the OHS, many of the key questions are asked the same way in each

[5] One explanation offered for the poor income data in the 1995 OHS is that Statistics South Africa knew they would be re-interviewing the same households a couple months later and so did not put a lot of time into getting complete responses to the income questions in October.

[6] In practice this means excluding individuals who age more than a year from October to December, whose reported gender or race changes, or who simply are not present in either one of the OHS or IES. We also are forced to exclude twins since it is not possible to identify which person is which when merging the IES and OHS.

of the surveys. This facilitates 1995-2000 comparisons. In the September 2000 LFS, there were about 105,000 individuals comprising about 26,600 households. The 2000 IES surveyed the same individuals. The format of the IES survey was the same in 2000 as in 1995 and questions were asked in identical ways. While we are sensitive to issues that arise in terms of how income information is gathered, the 1995 and 2000 Income and Expenditure Surveys are about as compatible as they could be.[7,8] After merging the 2000 IES and LFS, again being careful to delete individuals that do not match key demographics, we are left with just over 101,000 individuals comprising almost 26,000 households.[9]

When we then limit our analyses to individuals age 18 and older with valid information on age, race, gender, education, and sampling weight, we are left with 69,701 individuals in 1995 and 60,415 in 2000.[10] Appendix Table 3 lists the number of individuals in our sample in each year by age, race, and gender.

Making real comparisons requires adjusting for inflation. To do so, we employ CPI deflators provided by Statistics South Africa. The South African CPI data are available at the provincial level. Cross-province variation in price levels, though, is negligible, possibly reflecting the good infrastructure of South Africa relative to many of its African neighbors. The price data do not include prices from rural locations. In some countries, this would be a significant omission, but we do not believe this is the case in South Africa. Hoogeveen and Ozler (2004) report that their poverty and inequality analyses are unaffected by assuming that prices in rural areas are five percent higher. They also note that in the 2000 IES, approximately 60 percent of rural households report purchasing grain products in nearby urban areas. They also report that rural households bought most of their non-food items in urban areas. Finally, note that it is *changes* in the urban/rural relative price gap that will impact our analysis. There is little reason to believe these are significant. In the results reported, we use a single price deflater for the entire country.

[7] We elected not to use the 1993 LSMS survey for just this reason. The income questions in 1993 are not exactly compatible with those in the 2000 IES.

[8] Perhaps the only aspect of non-comparability in the data is the commonly-encountered difficulty of distinguishing zero values for income from missing values for income. For 1995, there are no missing values for total income. For 2000, just below 3% of individuals aged 18 and above have missing total income. For comparability, we code the 2000 missings as zeroes. This is not a correct procedure in terms of a cross-sectional analysis. However, it is precisely what we believe happened to the 1995 data, so is desirable in terms of preserving comparability across survey years. Note that this decision affects only our models for who receives income.

[9] We excluded individuals who aged more than a year from September to December, whose race and/or gender changed, or who simply are not present in either the LFS or IES.

[10] In 1995 and 2000, valid information on the variables listed was available for all but 739 and 772 observations, respectively.

3. A Descriptive Analysis

We begin with a simple descriptive analysis, highlighting the basic patterns guiding our investigations. Figure 1 presents kernel density estimates of log real individual income (in 2000 Rand) for 1995 and 2000 for individuals 18 and older.[11] This figure motivates almost everything that follows. The top panel gives the distributions for men while the bottom panel does so for women. For our purposes, there are three key aspects of Figure 1.

First and foremost, log real individual incomes declined between 1995 and 2000. For both men and women, the 2000 distribution lies clearly to the left of the 1995 distribution. With the exception of the very highest incomes, the shift is evident throughout the support of the distribution.

Second, incomes are generally lower for women than for men, and the decline in log real income is more severe for women than it is for men. In Figure 1, and throughout the paper, we report results for men and women separately. We do so because we find that males and females fare quite differently and so combining them often obscures more than it reveals. For example, women appear to have been joining the labor force during our period of study, while men appear to have been exiting the labor force.

Third, government old age pensions are a dominant form of income. These pensions are responsible for the large spike in each of the panels of Figure 1.[12] Pensions are especially important for women, as evidenced by the relatively larger mass near the typical pension income. Women become eligible for the old-age pensions at age 60 and men at age 65.

The data underlying Figure 1 are summarized in Table 1. This table presents more readily interpretable magnitudes of the decline in incomes throughout the distribution and presents separate information for wage and salary income only as well as pension income only.[13] Although South African incomes declined virtually throughout the distribution, the change is starkest in the lower half of the distribution. For example, for men, the 1st and 5th percentiles in 2000 total real income are both roughly one-half what they were in 1995; the 10th through the 75th are roughly one-third

[11] In all the results in this paper, we limit the sample to individuals 18 and older unless otherwise noted.

[12] For both men and women earning pension income, the median of log pension income in 1995 is 8.84; for both men and women, the spike in the density occurs at 8.86. In 2000, the median is 8.78 and the spike occurs at 8.80 (men) and 8.77 (women). Evidently, those earning pension income receive only small amounts of income from other sources. According to Woolard (2003), the "maximum grant payable per month as approved" held roughly constant in real terms from 1995 to 2000: 575 Rand in 1995 and 550 Rand in 2000, in 2000 prices. Observe that $\ln(575 \cdot 12) \doteq 8.84$ and $\ln(550 \cdot 12) \doteq 8.79$. The actual monthly grant payable may rise to roughly double the "maximum grant payable per month as approved", with additional income effectively taxed at a 50% rate (Woolard (2003, fn. 2)).

[13] That is, for the subset of persons with non-missing and positive wage and salary income and pension income, respectively. The column for all income includes these two as well as other transfers and remittances.

less than what they were in 1995; and above the 90th, incomes declined by about one-seventh. For women, the results are essentially the same, but above the 90th percentile, women fare slightly better than men, with roughly constant real incomes. Before proceeding, we stress that the decline in real incomes these data document is of a large magnitude in economic terms by almost any metric. Our key goal in this paper is to better understand this seismic shift.

When one examines the data that underlie the panels of Figure 1 as well as Table 1, it is striking that on average real incomes fall by about the same general magnitude as inflation. Put another way, one would not be terribly wrong if one posited that, on average, nominal wages and salaries were about fixed and inflation eroded their real values between 1995 and 2000. In real terms, the median wage and salary in 1995 and 2000 were R21,030 and R15,600 respectively. In nominal terms, the 1995 median was R15,000 while in 2000 it was R15,600. On first pass, a model of nominal wage rigidities would find support in the aggregated descriptive data.[14]

The general decline in real incomes from 1995 to 2000 evidenced by Figure 1 is a robust finding when using South African income and expenditure surveys. Hoogeveen and Ozler (2004) report comparable real declines in *household* incomes. In Appendix Figure 1, we re-visit their findings with our data and confirm that the distribution of household log real incomes fell from 1995 to 2000. In the bottom panel of that same figure, we also show that the distribution of real household expenditure also fell.

Although the survey data is unanimous in painting a dismal picture of income and expenditure declines, national income figures paint a brighter picture with real income either holding steady or even increasing.[15] We are silent on this glaring discrepancy except to note that: i) This type of discrepancy is not uncommon and is documented in other contexts and discussed in Deaton (2004); and ii) The results using the survey data are replicable and all the inputs—principally survey responses and sampling weights—are explicit and available hence yielding a relatively transparent and replicable methodology. The same cannot be said of the national incomes data.

In each panel of Appendix Figure 1, it is especially notable that the real household-level declines are apparent throughout the entire spectrum of income and expenditure levels. When examining

[14] In this regard, it is important to note that the data we obtained from DataFirst were not previously deflated. While we were assured that the data were in nominal terms and have no reason to disbelieve, skeptics will find comfort in the fact that the pattern of "rounding" in our incomes data is consistent with these assertions. Before deflating the 1995 data, roughly 88% of reported total individual incomes have a final digit of zero, rather than the expected 10%. After deflating, 11% of reported 1995 real incomes have a final digit of zero.

[15] Which case obtains depends on which national income statistic one examines. For example, real GDP at market prices as reported by StatsSA increases while real gross national income per capita as reported by the Reserve Bank is about constant.

household-level incomes and expenditures, zero incomes and/or expenditures are a non-issue—there are none. Hence, when examining household-level data, the selection issue does not arise, except to the extent that there is selection within the household. This would be the case, for example, if women within a household joined the labor market while men in the same household exited the labor market.

We next provide descriptive evidence on the endowments, returns, and selection explanations for the declines in real incomes. Each are discussed in turn. We conclude the section by investigating yet another possible explanation for the observed decline in real incomes—poor data integrity.

The Endowments Explanation

One candidate explanation for the decline in individual incomes is that individuals' endowments changed (for the worse). A decline in the level of education, for example, would be such a change in endowments.

Table 2 speaks to the endowments explanation. The top panel gives average characteristics of the male population by year for the entire sample as well as only for those reporting positive income. The bottom panel reports the analogous information for females. The overwhelming message from this table is rather simple—changes in endowments are small in terms of economic magnitudes.[16] The comparisons of mean characteristics conditioning on those with positive income highlights some expected patterns. Those with positive incomes tend to be older, whiter, and more urbanized. An unexpected finding is that those with positive incomes are not, on average, better educated.[17] Among those with positive incomes, mean age, education, and the percent urban are virtually unchanged from 1995 to 2000. Thus, Table 2 suggests that the endowments explanation is somewhat empirically suspect. There just does not appear to be the sort of large changes one would expect if the endowments explanation was driving the shift in the distribution of individual income. The figures in Table 2, though, are merely averages and as such are not dispositive when it comes to explaining changes in the entire distribution.

[16] Probably the sole exception to this is the increase in the fraction of the population reported as Black. This has led some to question the 2000 sampling weights. Anecdotally, many argue that systematic emigration of South African Whites may have been an important phenomenon between 1993 and 1995, but was largely over by 1995. On the other hand, as recently as April 21, 2005, *The Economist* reported that "It is likely, however, that 250,000-plus [Whites] have left since 1994, many of them young and talented." Our results in Table 6 are more consistent with *The Economist* than with those who argue that the emigration was mostly complete by 1995.

[17] Closer investigation reveals that this pattern is due to the secular improvement in education in South Africa. Younger individuals are less likely to receive income due to the youth unemployment problem, but are more educated than their older counterparts. For each separate age, those with positive income are more educated than those with no reported income (overall, as well as by survey year by gender).

The Returns Explanation

Another explanation for the decline in individual real incomes is that the returns to individuals' endowments changed in such a way as to lower income. To fix ideas, Table 2 showed that the mean level of education was virtually unchanged from 1995 to 2000. However, if the return to this education were to have fallen, this would contribute to the observed decline in incomes. Table 3 provides a first-pass examination of the returns explanation. That table lists the coefficients from standard OLS Mincer regressions of individual log income on age (and age squared), years of education, indicators for race[18] , and an indicator for pension eligibility. Separate regressions are run for men and for women. The table also lists the difference in coefficients and the standard error of that difference.[19] The estimated coefficients suggest that returns did indeed change from 1995 to 2000.

Looking first at men who reported positive incomes, the coefficients on all the included races increased. Blacks are the excluded group so one can interpret this as showing that the return to *not* being Black increased. Somewhat surprisingly, the return to being White increased almost twenty points *after* the White government stepped down and the African National Congress assumed power. The return to being Coloured also increased 20 points and this too is somewhat surprising since anecdotal evidence often suggests that Coloured workers are the ones being left behind. (The argument is that Whites inherited positions of privilege and that Blacks are the main beneficiaries of various affirmative action programs, hence leaving Coloureds in a vulnerable position. A counter-argument, also mostly anecdotal, is that Coloureds have benefited from better schools than have Blacks.) Both the increase in the return to being White and increase in the return to being Coloured are precisely estimated. The return to age also increased (at standard levels of statistical significance). The return to years of education was virtually unchanged while the return to being pension-eligible increased by about five points, although the difference is not precisely estimated.

For women, the pattern is similar. Again, the return to not being Black increased and the return to being White increased dramatically. Returns to age (experience) also increased. Returns to years of education and pension-eligibility were essentially unchanged.

[18] In order to measure progress towards the dismantling of apartheid, the post-apartheid government continues to monitor the four race demarcation that was inherited from apartheid. In 1995, approximately three quarters of the population were Black, just over 10 percent were White, just under 10 percent were Coloured, and about 3 percent were Asian/Indian. See UNDP (2003).

[19] See notes in Table 3 for details on how these standard errors were computed.

The Selection Explanation

Figure 1 showed the shifts in the distribution of log income. These figures do not include individuals who earned zero incomes. To the extent that the set of individuals who earn income might change over time, a selection issue arises. For example, if it were the case that in 2000 there were simply more Black workers (who tend to earn lower wages) and more poorly educated workers among the employed, these changes alone could in principle explain the observed shift in individual log incomes (at least for the left-hand-side of the distribution). More generally, selection into income recipiency may not be random and the likelihood of a particular individual earning positive income may change between the sample periods. In this subsection, we provide a first pass investigation of the selection explanation.

Table 4 reports income recipiency rates (the fraction of the population that reported strictly positive income) for several different subpopulations. Although we will explore the selection explanation in greater detail in our nonparametric analysis, this table illustrates some patterns in the data that those analyses must confront. For example, overall income recipiency rates are essentially equal in the two periods. The overall fraction of the sample reporting positive income was 58.0 percent in 1995 and it was relatively unchanged at 56.7 percent in 2000. This finding is corroborated by the unemployment data in our data sets. Although zero incomes and unemployment are not one and the same, our data show relatively little change in unemployment from 1995 to 2000.

However, gross income recipiency rates mask stark differences by gender. For example, men in 2000 are, on average, less likely to receive any income than in 1995, while women are more likely to receive income in 2000 than in 1995. This gender difference is a dominant pattern in the data, evidently more important than race, age, or education. For example, the difference-in-differences of women's income recipiency rate changes versus men's income recipiency rate changes is over 10 percentage points. This difference-in-differences is essentially the same for all four population groups and all but the youngest age ranges (where many individuals may yet be in school).

As with any descriptive analysis, the simple differences in income recipiency in Table 4 are subject to multiple interpretations. For example, those with high levels of education are more likely to receive income, but this could in principle be due to education itself, or to the correlation of education with race. Conversely, education is predictive of youth, which is negatively related to income recipiency due to youth unemployment, so differences by educational attainment may understate the effect of education.

Some of these hypotheses may be ruled out by the more disaggregated analysis of Appendix Table 1, which reports income recipiency rates for 1995 and 2000 by cells defined by age, race,

sex, and education. The disaggregated rates reveal several interesting patterns. First, for 48 of the 54 cells for men, income recipiency rates fell between 1995 and 2000. Income recipiency rates fell precipitously for White men.

Second, and conversely, for 34 of the 54 cells for women, income recipiency rates rose. In particular, Black women with twelve or fewer years of education uniformly saw improvements in income recipiency rates between 1995 and 2000. For non-Black women, income recipiency rates were roughly statistically equivalent.

Third, on a cell-by-cell basis, education is protective against non-recipiency of income. However, for men, the extent to which education is protective appears to be roughly equal in 1995 and 2000. For women, education was much less protective in 2000 than in 1995, particularly for Black women.

A more parsimonious analysis is presented in Table 5, which gives marginal effects for logit models for income recipiency in 1995 and 2000, separately by gender. Marginal effects are calculated as the sample average of the derivative of the conditional probability of receiving income.[20] Turning to the coefficients, we see that in 1995, White men were 28.6 percentage points more likely than Black men (the excluded category) to receive a positive income. By 2000, that figure had declined to 13.9 percent. The conditional "advantage" to being White declined almost 15 percentage points—a decline that was both large and precisely estimated. Similarly, the advantage enjoyed by Indian men in terms of income recipiency also declined relative to Black men. For women, patterns are similar but more muted.

The conditional impact of pension eligibility is essentially unchanged across years. This suggests that differential take-up rates for pensions in 1995 and 2000 are not important when considering the selection explanation.

The fact that so many of the differences in Table 5 are non-zero (and precisely so) buttresses the key message from Table 4—selection may be important. For example, suppose Black women's income recipiency rates rose. Suppose these women earn relatively low wages. Their inclusion in the distribution of non-zero incomes will tend to give more mass to the lower tail even if endowments and returns to those endowments remained constant. On the other hand, Figure 2 suggests that

[20] That is, if $\hat{\beta}$ are the estimated logit coefficients, and $P(D_i = 1|X_i) = F(X_i'\beta)$ where $F(t) = e^t/(1+e^t)$, then the marginal effects reported are $\sum_{i=1}^{n} V_i f(X_i'\hat{\beta})\hat{\beta}$, where $f(t) \equiv F'(t) = e^t/(1+e^t)^2$, and V_i is a sampling weight that sums to one. Standard errors for the marginal effects, reported in parentheses, are calculated by the delta-method. Specifically, define $g(\beta) = \sum_{i=1}^{n} V_i f(X_i'\beta)\beta$ and note that $Dg(\beta) = \sum_{i=1}^{n} V_i f(X_i'\beta)I_k + \sum_{i=1}^{n} V_i f'(X_i'\beta)X_i\beta'$ is the derivative of $g(\cdot)$ with respect to β, where $f'(t) = e^t(1-e^t)/(1+e^t)^3$ and k is the number of parameters to be estimated. Then the square-root of the diagonal of $\hat{V}[g(\hat{\beta})] = Dg(\hat{\beta})'\hat{V}[\hat{\beta}]Dg(\hat{\beta})$ is a consistent estimate of the standard deviation of the marginal effects, where $\hat{V}[\hat{\beta}]$ is the estimated variance matrix for the logit coefficients, calculated using the so-called "cluster" option at the level of the household.

if selection is at work, then it must be of a relatively complicated variety. This figure gives age profiles for mean log total income separately for men and women for 1995 and 2000. Despite their very different patterns in regards to income recipiency over time, the experience of men and women have been remarkably similar in terms of the erosion of real income. Roughly speaking, men and women experienced equal magnitudes of income decline, with young people bearing the brunt of the burden. We reserve a more careful analysis for Section 4.

The "Bad Data" Explanation

There is, of course, another explanation for the shift in the distribution of income and that is that the data are somehow contaminated in a way that makes 1995 versus 2000 comparisons simply invalid. This is always a lurking suspicion when using repeated cross-sections of data. These concerns, as noted above, are alleviated due to the similarity of the survey instruments over time and the fact that the same government organization (Statistics South Africa) conducted both surveys. Still, it is worthwhile conducting some analyses to investigate the integrity of the data. Toward that end, Table 6 examines means of education and race for given cohorts.

Abstracting from emigration, these means should be similar in 1995 and 2000, because race is immutable and most individuals are no longer enrolled in school. For example, in the first row of Table 6 we compare the mean of the indicator variable for White for the cohort born between 1931 and 1935. The reported difference is in fact zero. For two of the male cohorts, there are statistically significant declines in the fraction of the population that is White although these declines are modest. Anecdotal evidence suggests that some small declines are perhaps to be expected since there has probably been some emigration by Whites. There are also modest but statistically precise declines in White females in these same cohorts as well as two others. There are no statistically significant changes for any of the male or female Coloured and Indian cohorts.

The bottom panel computes mean education levels by gender and cohort. For those cohorts in the middle of the age distribution, there are statistically significant declines in the average number of years of education for both men and women while for the others there are no such changes. Although the exact magnitude of the differences vary by gender and cohort, when the differences are precisely estimated they are on average about a half a year. With a return to an extra year education at a bit below twelve percent (cf. Table 3), a back-of-the-envelope calculation suggests that the different educational endowments could explain around a six percent income decline for about half of the cohorts (and not the cohorts with the largest numbers of people). While the

declines in reported education levels for some cohorts give one pause, they are too small to explain the large income declines actually observed throughout the income distribution.

Finally, it is important to realize that even if the surveys were implemented to perfection, one would not expect to see all zeroes in the column of differences. The two data sets are not a panel so different individuals comprise each cohort. While conventional wisdom suggests emigration had substantially slowed by 1995, it may not have been complete by then, and it is (presumptively) not random across race and education. Also, differential mortality rates between 1995 and 2000 may contribute to some differences. In sum, we conclude that the income and expenditure data and accompanying demographic data are reliable.

Even if the income and expenditure data are spot-on, it could be the case that the price index data are corrupted. Mechanically, mis-measured price indexes will generate mis-measured real income changes. We use price index data provided to us directly by Statistics South Africa.[21] Although we use an aggregate CPI to adjust incomes, we have obtained much of the disaggregated price data.[22] The disaggregated price data is itself an average of actual price quotes collected by Statistics South Africa.[23] The price index data we use matches the publicly available price index data available from both Statistics South Africa and from the Reserve Bank of South Africa. This is not surprising since the data provided to us by Statistics South Africa are the same data used to compile the publicly available CPI.

We have also sought evidence that might corroborate our finding of declining individual log incomes. Figure 3 examines what has happened to the share of food expenditure as a fraction of total household expenditure from 1995 to 2000.[24,25] An attractive feature of Figure 3 is that it does *not* depend on the price deflators since each observation is the share of food in total expenditure for a given household in a given year. The shift in the share of food expenditure is dramatic.[26] From

[21] The data were provided at the request of the Office of the Presidency as Leibbrandt and Levinsohn were preparing a report for that office.

[22] For 1995 and 1996, no data were stored electronically and as of May 2003, the computer system used for compilation of the CPI could only store 24 months of data. This information as well as a description of how the disaggregated price data are used to create the CPI are from a personal letter dated May 3, 2003 from P.J. Lehohla, the Statistician General of StatsSA to one of the authors.

[23] For example, for mealie meal, approximately 30 price quotes are obtained for a given province.

[24] We are grateful to Jere Behrman for suggesting this figure.

[25] A more detailed examination of food expenditure on a per capita basis for South Africa (and several other countries) can be found in Deaton and Paxson (1998). That study uses a data source, the 1993 LSMS survey, that we do not use because questions are asked in a very different way making comparisons with 2000 problematic.

[26] For comparison purposes, the food expenditure questions for 1995 and 2000 are identical with one minor excep-

1995 to 2000, the distribution of the share of household expenditure devoted to food shifts sharply to the right. In 1995, the distribution peaks at around 20 percent. While there is a remnant of that peak in 2000, the distribution flattens out. The distributions cross at about 35 percent. After 2000, there are more households spending over 35 percent of expenditure on food and few spending less compared to 1995. This is entirely consistent with a substantial decline in real income, and this figure does not depend on either income or price deflater data.

In sum, we stand behind the quality of the data that we analyze, particularly in regards to our general conclusions. We seriously doubt, for example, that the decline in real incomes we document here is an artifact of faulty sampling or mistakes in processing. Essentially, the data we analyze give every indication of being of the caliber of the standard repeated cross-section household survey conducted in any developed country.

Figure 1 set the stage. It cleanly illustrated the motivation for the rest of the paper. Table 2 then provided descriptive evidence suggesting that the endowments explanation is unlikely to be important. Table 3 examined changes in returns in an exceedingly simple way. This first-pass reinforces the possibility that changes in returns might, at least in part, be behind the evidence in Figure 1. Tables 4 and 5 lend some credence to the selection explanation. Finally, Table 6 and Figure 3 suggest that simply writing off the decline in incomes as an artifact of the data is presumptive (and perhaps wrong). We now turn to a more powerful approach to investigating the endowments, returns, and selection explanations.

4. A Nonparametric Approach

In this section, we apply relatively simple nonparametric techniques to investigate the endowments, returns, and selection explanations for the shift in the distribution of individual log real incomes displayed in Figure 1. Each possible explanation is analyzed in turn. For each explanation, we first describe our methodology in general terms. We then present an algorithm or "recipe" to explain just how to implement the methodology. This may be helpful since while some of the methods may appear daunting, they are in practice pretty simple.[27] We then discuss results for the particular explanation under consideration.

---

tion. In 2000, expenditure on baby food was a separate section of the survey instrument. These questions, though, were present and identical in the 1995 survey instrument also, but in 1995, they did not comprise a separate section of the food expenditure survey. That is, the difference is purely cosmetic.

[27] Some of the nonparametric techniques we use were developed in the course of working on the substantive question that is the topic of this paper. In a companion paper, Levinsohn and McCrary (2005), two of us report technical details, give computer code for implementing the estimators we describe here, and report results from Monte Carlo simulations.

Before proceeding, we note that the counterfactual questions we pose and estimate in this section are similar to those explored by several different authors. Autor, Katz, and Kearney (2004) discuss a quantile regression approach to estimating the counterfactuals discussed in Juhn, Murphy, and Pierce (1993), building on the quantile regression methodology of Machado and Mata (2005). The reweighting approach that inspires our own methodology was described in DFL and further amplified by Lemieux (2002).

The endowments explanation: Methodology

To investigate the role that changes in endowments might have played in shifting the distribution of log real incomes, we straightforwardly apply the approach of DFL. We start by setting notation.

The density functions for income in periods t and t' may be written as

$$f(y|T=t) = \int g(y|x, T=t)h(x|T=t)dx \tag{1}$$

and

$$f(y|T=t') = \int g(y|x, T=t')h(x|T=t')dx \tag{2}$$

respectively, where T is a random variable describing the year from which a given individual in the *pooled* dataset of observations from both survey years is drawn, $g(y|x, T=t)$ is the density of individual income evaluated at y, given that the observable attributes[28] of the individual, X, are equal to x and that the survey year is t, and $h(x|T=t)$ is the density of attributes evaluated at x, given that the survey year is t. It is perhaps helpful to think of $g(y|x, T=t)$ as the function that "translates" observable attributes into income. Were this a traditional parametric regression of income on individual endowments for a given year t, the density of individual income, $f(y|T=t)$, would be analogous to the dependent variable, income; $h(x|T=t)$ would be analogous to the endowments data; and $g(y|x, T=t)$ would be analogous to the returns to those endowments.

We will be interested in how the density of individual income changes if attributes and/or returns to those attributes changed. To do so, it is necessary to define precisely what we mean by these "counter-factuals". Suppose that we are interested in how the distribution of income in period t would differ, were the *endowments* as they were in period t'. That is, what if individuals' endowments were those that obtained in 2000 (t') instead of the actual 1995 (t) endowments? We denote this counter-factual by $f_h^{t\to t'}$; it may be written symbolically as

$$f_h^{t\to t'}(y) \equiv \int g(y|x, T=t)h(x|T=t')dx. \tag{3}$$

[28] Throughout the paper, we refer interchangeably to "observable attributes", "attributes", "endowments", and even "observables".

Notationally, the subscript "h" indicates that it is the density of attributes, or $h(x|T=t)$, that is being changed from an actual to a counter-factual density. The superscript, "$t \rightarrow t'$" indicates that in this counter-factual, we are going to start with *data* from period t and use statistical techniques, in particular a re-weighting scheme, to transform the actual density of attributes from the $h(x|T=t)$ that reigned in period t to the counterfactual density $h(x|T=t')$ that reigned in period t'.[29]

Two specific examples help to further fix ideas. If we wanted to ask what the 2000 (t') distribution of income would look like if instead endowments were distributed as they were in 1995 (t), we would estimate the counter-factual density

$$f_h^{t' \rightarrow t}(y) \equiv \int g(y|x, T=t')h(x|T=t)dx. \tag{4}$$

This would involve using data from 2000 and using a re-weighting scheme to make the density of endowments "look like" the density of endowments in 1995. If we wanted to ask what the 1995 distribution of individual income would look like if *returns* to endowments were as they were in 2000, we would estimate:

$$f_g^{t \rightarrow t'}(y) \equiv \int g(y|x, T=t')h(x|T=t)dx. \tag{5}$$

Comparing equations (4) and (5), we see that the notation conventionally used to describe counterfactuals is somewhat inspecific; the right-hand-sides of (4) and (5) are identical, but $f_h^{t' \rightarrow t}$ is based on data from t' while $f_g^{t \rightarrow t'}$ is based on data from t. Moreover, these two distributional parameters pertain to counterfactuals which are conceptually entirely distinct.[30]

The fundamental insight from DFL is that the counter-factual in (3) is easy to implement using a re-weighting idea. We first describe how this is done and then show how this insight can be extended straightforwardly to the other counterfactuals. The re-weighting idea of DFL is based on the simple recognition that Bayes' Axiom implies

$$\frac{h(x|T=t')}{h(x|T=t)} = \frac{P(T=t'|X=x)}{1-P(T=t'|X=x)} \Big/ \frac{P(T=t')}{1-P(T=t')} \equiv \tau_h^{t \rightarrow t'}(x) \tag{6}$$

[29] As we describe in more detail below, the notation is modular. For example, $f_h^{t' \rightarrow t}(y)$ would indicate a counterfactual density one would estimate starting with data from period t' and using a re-weighting scheme to transform the density of attributes from $h(x|T=t')$ (the actual attribute density) to $h(x|T=t)$ (the counter-factual attribute density).

[30] This reinterpretation immediately suggests that we could also view the the right-hand-side of equation (2) as addressing the counter-factual: "What if individuals in 1995 had the endowments from 2000 instead of the actual 1995 endowments, *and* enjoyed the returns to endowments that obtained in 2000, instead of the actual 1995 returns?" We describe below how to compute this counterfactual as well.

In words, $\tau_h^{t\to t'}(x)$ is just the ratio of the conditional odds to the unconditional odds.[31]

This turns out to be precisely the weighting function needed to conduct the endowments counter-factual of (3). To see this, rewrite the object of interest $f_h^{t\to t'}(y)$ as

$$\begin{aligned} f_h^{t\to t'}(y) &= \int g(y|x,T=t)h(x|T=t')dx = \int g(y|x,T=t)h(x|T=t)\frac{h(x|T=t')}{h(x|T=t)}dx \\ &= \int g(y|x,T=t)h(x|T=t)\tau_h^{t\to t'}(x)dx \end{aligned} \tag{7}$$

which differs from (1) only by the weight $\tau_h^{t\to t'}(x)$. Consequently, if we could estimate the weighting function $\tau_h^{t\to t'}(x)$ then we could compute the counter-factual (3) easily using a *weighted* density estimate of incomes (with a density estimation technique of our choosing).

Considering the structure of the weighting function, estimation of the counter-factual is quite straightforward.[32] Predicted probabilities $\widehat{P}(T_i = t'|X_i)$ may be obtained from a binary choice model (such as a binary logit) that uses individual attributes X_i to predict the probability of an observation coming from year t' in the pooled data set of observations from both years. The predicted probability $\hat{P}(T_i = t')$ may be obtained from the relative frequency, and we may reweight the data using a plug-in version of (6). This procedure is summarized in the following algorithm:

[31] The odds of an event that occurs with probability p is $p/(1-p)$. The expression in the display is not to be confused with the "odds ratio" which is the ratio of the odds for the treatment group to the odds for the control group in the randomized control trial.

[32] The ease with which DFL can in fact be implemented may not be obvious to most readers of the original *Econometrica* paper, although the description in Johnston and DiNardo (1996) is admirably simple. To facilitate replication, all of the code used in our paper will be available at `http://www.umich.edu/~jmccrary`. All estimation is programmed in Stata version 8.2.

Endowments Algorithm

1. Using the *pooled* dataset of observations from both survey years, estimate the fraction of observations with $T_i = t'$, or $\hat{P}(T_i = t')$.
2. Using the pooled dataset, estimate a logit for $T_i = t'$ using individual attributes, or endowments, X_i. Store the predicted values from the logit, or $\hat{P}(T_i = t'|X_i)$.
3. For the *subsample* of observations from year t, generate a new weight

$$\hat{\tau}_{i,h}^{t \to t'} \equiv \frac{\hat{P}(T_i = t'|X_i)}{1 - \hat{P}(T_i = t'|X_i)} \Big/ \frac{\hat{P}(T_i = t')}{1 - \hat{P}(T_i = t')}$$

4. For the subsample of observations from year t, use a weighted kernel density routine to estimate

$$\hat{f}_h^{t \to t'}(y) = \frac{1}{b} \sum_{i=1}^{n} K((Y_i - y)/b)\, V_i\, \hat{\tau}_{i,h}^{t \to t'}$$

   where Y_i is log income, $K(\cdot)$ is a kernel function, b is a bandwidth parameter, and V_i is the sampling weight. If the survey is self-weighted, then $V_i = \frac{1}{n}$, where n is the number of observations from year t. If the sample is not self-weighting, then one should verify that the estimated probabilities from steps 1 and 2 employ the sampling weight.

For comparison, note that the conventional density estimate for actual log incomes in year t is

$$\hat{f}(y|T_i = t) = \frac{1}{b} \sum_{i=1}^{n} K((Y_i - y)/b)\, V_i.$$

This makes it clear that estimating the counterfactual density function is no more challenging than estimating a density function, given an estimate of the weighting function.[33]

Suppose we were interested in a very specific counter-factual pertaining to endowments, such as "What would the distribution of log income in 2000 look like, were the educational endowment of the population as it was in 1995?" To answer this question, we interpret (3) as an integral over the distribution of *education alone*. Then we apply the Endowments Algorithm, with Step 2 being a logit model for $1(T_i = t')$ using education (X_i, a scalar) as a predictor—and possibly higher powers of education, depending on how nonlinear in X_i one believed the true conditional probability $P(T_i = t'|X_i)$ to be.

We in fact focus on a broader counter-factual: "What would the distribution of log income in 2000 look like, were all endowments to be as they were in 1995?" To answer this question, we

[33] It is conventional to assess the counter-factual by comparing $\hat{f}_h^{t \to t'}(y)$ to $\hat{f}(y|T_i = t)$ and $\hat{f}(y|T_i = t')$. While it turns out that it makes little difference in our application, we believe that the "returns and endowments" counter-factual described below is a slightly better basis for comparison than $\hat{f}(y|T_i = t')$.

interpret (3) as an integral over the distribution of *all observables*—education, race, and so on. Step 2 of the Endowments Algorithm is then a logit model for $1(T_i = t')$ using all observables (X_i, a vector)—and possibly powers and cross-products of the elements of X_i.[34]

The endowments explanation: Results

Implementing the above methodology requires specifying a measure of income, Y_i, as well as the components of the vector of individual attributes, X_i. In the notation above, t is 1995 while t' is 2000. Our measure of individual income is the log of total individual income. It includes wage income as well as income from all other sources (such as pension income).[35] The vector of individual attributes is comprised of dummies for whether the individual is Black, Indian, or Coloured (with White the excluded group), age, age squared, a dummy that takes on the value of one if the individual is male and over 65 or if the individual is female and over 60 (to capture pension eligibility), and years of education. These are our measures of an individual's endowments.

An intuitive way to examine the results is to compare the estimated distribution $\hat{f}_h^{t \to t'}(y)$ to the actual 1995 distribution. In words, the estimated distribution is illustrating what the distribution of income would have been in 1995 if endowments were those that obtained in 2000 but all else was the same.

The results are given in Figure 4. The top panel is for the male subsample while the bottom panel is for the female subsample. These diagrams are easy to interpret. For both men and women, the reweighted distribution essentially lies coincident with the 1995 actual distribution. For men, the upper tail of the density estimates diverges slightly, but the differences are quite minor compared to the important differences between the 1995 and 2000 density estimates of Figure 1. Consequently, speaking loosely, we might say that substituting 2000 endowments for 1995 endowments gives rise to the 1995 actual distribution of income. We interpret this as convincing evidence that the endowments explanation makes virtually no headway in explaining the shift in individual incomes. This is broadly consistent with the descriptive analysis. Two additional comments are relevant. First, recall that the descriptive analysis showed an increase in the fraction of the population that was Black. Because Black South Africans earn less, one might have imagined that this simple demographic difference could have possibly explained the decline in incomes. Indeed, in the logit

[34] We do so because we find little evidence that any individual endowment explains the decline in real incomes. The focus on the omnibus endowments counter-factual is thus a summary of this disaggregated evidence.

[35] In many developing countries, income in the form of self-production of food is an important component of income and exactly how this is modeled becomes important. In South Africa, income attributed to self-production is negligible. It is not included in our definition of individual income since it is only reported at the household-level.

model for $P(T_i = 1|X_i)$, race is highly predictive (c.f. Appendix Table 2). However, the results of Figure 4 indicate that the Black income penalty is not nearly large enough to account for the overall decline in real incomes. Second, if it were the case that the explanatory variables that comprise X_i in fact don't "explain" income, then the results in Figure 4 would be non-informative. However, the results in Table 3, discussed in the descriptive analysis section, above, indicate that this is not the case.

The returns explanation: Methodology

Here, we are interested in how the distribution of income in period t would differ, were the *returns* to observables as they were in period t'. Following the notation of above, we label this counter-factual by $f_g^{t\to t'}$ and note that it may be written symbolically as

$$f_g^{t\to t'}(y) \equiv \int g(y|x, T = t')h(x|T = t)dx \tag{8}$$

It turns out that this counter-factual, too, is easy to implement using a re-weighting idea. We again use Bayes' Axiom to derive an appropriate weight

$$\frac{g(y|x, T = t')}{g(y|x, T = t)} = \frac{P(T = t'|X = x, Y = y)}{1 - P(T = t'|X = x, Y = y)} \Big/ \frac{P(T = t'|X = x)}{1 - P(T = t'|X = x)} \equiv \tau_g^{t\to t'}(x, y) \tag{9}$$

and then note that the object of interest may be rewritten

$$\begin{aligned} f_g^{t\to t'}(y) &= \int g(y|x, T = t')h(x|T = t)dx = \int g(y|x, T = t)h(x|T = t)\frac{g(y|x, T = t')}{g(y|x, T = t)}dx \\ &= \int g(y|x, T = t)h(x|T = t)\tau_g^{t\to t'}(x, y)dx \end{aligned} \tag{10}$$

as before. Estimation of $\tau_g^{t\to t'}(x, y)$ is no more difficult than estimation of $\tau_h^{t\to t'}(x)$, as the following algorithm makes clear:

Returns Algorithm

1. Using the *pooled* dataset, estimate a logit for $T_i = t'$ using X_i. Store the predicted values from the logit, or $\hat{P}(T_i = t'|X_i)$.
2. Using the pooled dataset, estimate a logit for $T_i = t'$ using X_i *and* log income Y_i. Store the predicted values from the logit, or $\hat{P}(T_i = t'|X_i, Y_i)$.
3. For the *subsample* of observations from year t, generate a new weight

$$\hat{\tau}_{i,g}^{t \to t'} \equiv \frac{\hat{P}(T_i = t'|X_i, Y_i)}{1 - \hat{P}(T_i = t'|X_i, Y_i)} \Big/ \frac{\hat{P}(T_i = t'|X_i)}{1 - \hat{P}(T_i = t'|X_i)}$$

4. For the subsample of observations from year t, use a kernel density routine to estimate

$$\hat{f}_g^{t \to t'}(y) = \frac{1}{b} \sum_{i=1}^{n} K((Y_i - y)/b)\, V_i\, \hat{\tau}_{i,g}^{t \to t'}$$

   where the notation is as described in the Endowments Algorithm, and the cautions regarding the sampling weight V_i continue to apply.

We focus on the broad counter-factual—"What would the distribution of log income in 2000 look like, were the returns to all observables as they were in 1995?" To answer this question, we interpret (8) as an integral over the distribution of all observables. Step 1 of the Returns Algorithm is then a logit model for $1(T_i = t')$ using all observables (X_i, a vector)—and possibly powers and cross-products of the elements of X_i. Step 2 is the same logit model but with Y_i included—and possibly interactions of Y_i with X_i and the powers and cross-products of elements of X_i.

As noted in the above discussion of the Endowments Algorithm, it is straightforward to consider changes in but one or even a subset of the endowments. For example, we might want to know what the 1995 distribution of income would look like if the age profile was that of 2000. The thought experiment in this instance is well-defined, because both age and education are observable, and it is possible to imagine altering an individual's education without altering the individual's age.

The same is not true of changes in the returns to endowments. For example, one might wonder what would have happen to the 1995 distribution of income, were the return to education to be that which obtained in 2000 and other returns were unchanged—however, in all but one case, this thought experiment is flawed. If income is an additively separable function of each individual endowment, then it is straightforward to conceptualize changing only one return while holding the others constant. However, if the mapping from endowments to income is more complex, then the situation is more subtle.

A simple example illustrates the point. Suppose income (Y) is a function of only age (A) and

schooling (S) and the conditional expectation of Y given A and S is not quite additively separable:

$$E[Y|A,S] = \beta_1 A + \beta_2 S + \beta_3 A \cdot S$$

The return to a factor is the partial derivative of income with respect to that factor. Hence the returns to age and schooling, respectively, are given by:

$$\begin{aligned}\frac{\partial E[Y|A,S]}{\partial A} &= \beta_1 + \beta_3 S \\ \frac{\partial E[Y|A,S]}{\partial S} &= \beta_2 + \beta_3 A\end{aligned}$$

In this case—which we emphasize is only a modest departure from additive separability—the thought experiment of changing the return to age but not that of education makes sense only if we restrict β_3 to be the same across time periods. If changing the return to education involves changing β_3, then that thought experiment simultaneously changes the return to age. The problem illustrated in this example is quite general and indicates that the idea of changing returns "one at a time" is essentially only well-posed given very specific parametric models.

The leading case in which one *can* decompose the returns explanation on a return-by-return basis was already discussed in Section 3 (Table 3). There, we saw that the return to education for males was about 0.11 in 1995 and 2000. The same result applied to women. These aggregated coefficients mask considerable variation across age, race, and gender.

Disaggregated returns to education are given in Appendix Table 3. Separate regressions are run for each of six age cohorts for each race for each gender in each year. The table also lists the difference in estimates for each cell, whether that difference is statistically significant at the 5 percent level, and the number of observations in each cell. The first message of Appendix Table 3 is the dispersion of estimates across cells. This indicates that the data essentially reject the simplest versions of additive separability at a cross-sectional level. Second, there are some important patterns in the changes over time. This specifically rules out the validity of returns counterfactuals conducted "one at a time". In general, returns to education for the younger cohorts of Black men and women fell between 1995 and 2000. The declines were substantial—on the order of 4 or 5 percentage points (i.e., a decline from a return of 14% for an extra year of education to a return of 9%). At the same time, there were substantial increases for these same younger cohorts of White men and women. The fact that the changes in returns are especially evident for the younger cohorts could be taken to suggest that these differences over time may be picking up the impact of new hires or fires (in a system in which seniority matters).

The returns explanation: Results

We turn now to estimating the distribution of income that results when all the 2000 returns are applied to the 1995 data. The results of the logit regressions needed for steps 1 and 2 of the "Returns Algorithm" are reported in Appendix Table 2. As can be seen in the top panel of this table, other than income, the included regressors are the same as those used in the endowments explanation. We estimate (10) separately for men and women and report the resulting distributions in the panels of Figure 5. The solid line gives the actual 1995 distribution while the open circles give the reweighted 1995 distribution. Looking over the reweighted density, it appears that the "omnibus" returns counterfactual is nearly sufficient for explaining the discrepancy between the 1995 and 2000 (2000 actual density not shown; cf. Figure 1). However, as discussed above, it does not appear possible to tailor this approach to better understand which returns might have changed.

When trying to understand which returns changed and why, Figure 5 is usefully complemented by Appendix Table 3. Taken together, there seem to be two key patterns underlying the changes in returns—changes which themselves go a long way toward explaining the shift in the distribution of individual incomes. First, for Whites and Coloureds, the return to skill for those entering the labor market (ages 18-25) rose sharply. For Blacks, these returns fell. Falling average returns to education for Blacks and rising average returns to education for Whites have been noted before by Keswell (2004). Indeed, Keswell suggests that, controlling for age, the difference in returns had moved from close to zero in 1993 to close to 40 percent in 2002. Second, for Whites and Coloureds, returns for those in the other age cohorts either rose or stayed about constant while for Blacks, these returns either fall or stayed about constant.

For these findings to emerge out of the first five years following apartheid in South Africa is perhaps surprising, at least at first glance. We discuss this more in Section 5 below.

Considering the endowments and returns explanations together: Methodology

Finally, suppose that we are interested in how the distribution of income in period t would differ, were the observables *and* the returns to observables as they were in period t'. We label this counterfactual by $f_{g,h}^{t\to t'}$ and write it symbolically as

$$\begin{aligned} f_{g,h}^{t\to t'}(y) &= \int g(y|x,T=t')h(x|T=t')dx = \int g(y|x,T=t)h(x|T=t')\frac{g(y|x,T=t')}{g(y|x,T=t)}dx \\ &= \int g(y|x,T=t)h(x|T=t')\tau_g^{t\to t'}(x,y)dx \\ &= \int g(y|x,T=t)h(x|T=t)\tau_g^{t\to t'}(x,y)\frac{h(x|T=t')}{h(x|T=t)}dx \\ &= \int g(y|x,T=t)h(x|T=t)\tau_g^{t\to t'}(x,y)\tau_h^{t\to t'}(x)dx \\ &\equiv \int g(y|x,T=t)h(x|T=t)\tau_{g,h}^{t\to t'}(x,y)dx \end{aligned} \tag{11}$$

where $\tau_{g,h}^{t\to t'}(x,y)$ is defined by

$$\begin{aligned} \tau_{g,h}^{t\to t'}(x,y) \equiv \tau_g^{t\to t'}(x,y)\tau_h^{t\to t'}(x) &= \frac{\frac{P(T=t'|X=x,Y=y)}{1-P(T=t'|X=x,Y=y)}}{\frac{P(T=t'|X=x)}{1-P(T=t'|X=x)}}\frac{\frac{P(T=t'|X=x)}{1-P(T=t'|X=x)}}{\frac{P(T=t')}{1-P(T=t')}} \\ &= \frac{P(T=t'|X=x,Y=y)}{1-P(T=t'|X=x,Y=y)}\Big/\frac{P(T=t')}{1-P(T=t')} \end{aligned} \tag{12}$$

Because $\tau_{g,h}^{t\to t'}(x,y)$ only depends on the predictions $P(T=t'|X=x,Y=y)$ and $P(T=t')$, we may estimate $\hat{f}_{g,h}^{t\to t'}(y|T_i=t)$ using conventional density estimation techniques after estimating one binary choice model (to obtain the numerator in (12) and a relative frequency (to obtain the denominator).

When we simultaneously consider the endowments and the returns explanations, we are asking if it is possible to re-weight the 1995 data in such a way (allowing returns and endowments to change) so as to re-create the 2000 distribution. To the extent that the re-weighted distribution does not match the actual 2000 distribution, the most straightforward interpretation is that the support restrictions implicit in DFL are violated. That is, it may not be possible to re-weight data from one period or one group and obtain the equivalent distribution as in another time period or group. For example, it is conventional in studies of the impact of unions on wages *not* to attempt to re-weight the distribution of unionized workers to obtain counterfactuals pertinent to non-unionized workers. This is because there are, for example, few unionized workers with law degrees, but many non-unionized workers with law degrees. Rather, it is conventional to attempt to re-weight the distribution of non-unionized workers to obtain counterfactuals pertinent to unionized workers. This is because many non-unionized workers "look like they should be unionized".

We argue that a good interpretation of $f_{g,h}^{t\to t'}(y)$ is that this is "the best that the data from period t can do" in terms of replicating the distribution that reigns in period t'. We also note that this counterfactual allows for an interesting decomposition. We may write

$$\begin{aligned} f_{g,h}^{t\to t'}(y) &\equiv \int g(y|x,T=t)h(x|T=t)\tau_{g,h}^{t\to t'}(x,y)dx \\ &= \int g(y|x,T=t)h(x|T=t)\Big(\tau_h^{t\to t'}(x)+\Big(\tau_{g,h}^{t\to t'}(x,y)-\tau_h^{t\to t'}(x)\Big)\Big)dx \\ &= f_h^{t\to t'}(y)+\int g(y|x,T=t)h(x|T=t)\underbrace{\tau_h^{t\to t'}(x)\Big(\tau_g^{t\to t'}(x,y)-1\Big)}_{\equiv\,\tau_\delta^{t\to t'}(x,y)}dx \end{aligned} \tag{13}$$

where $\tau_\delta^{t\to t'}(x,y)$ simplifies to

$$\begin{aligned} \tau_\delta^{t\to t'}(x,y) &= \frac{\frac{P(T=t'|X=x)}{1-P(T=t'|X=x)}\,\frac{P(T=t'|X=x,Y=y)}{1-P(T=t'|X=x,Y=y)}-\frac{P(T=t'|X=x)}{1-P(T=t'|X=x)}}{\frac{P(T=t')}{1-P(T=t')}} \\ &= \Big(\frac{P(T=t'|X=x,Y=y)}{1-P(T=t'|X=x,Y=y)}-\frac{P(T=t'|X=x)}{1-P(T=t'|X=x)}\Big)\Big/\frac{P(T=t')}{1-P(T=t')} \end{aligned} \tag{14}$$

which is proportional to the gain in the prediction that $T=t'$ attributable to the event $Y=y$. The empirical version of this decomposition is similar. We note that this provides a formal justification for a procedure similar to, but distinct from, that pursued by DFL.

Considering the endowments and returns explanations together: Results

Figure 6 reweights the 1995 actual distribution using $\tau_{g,h}^{t\to t'}(x,y)$ as weights, allowing both endowments and returns to change. Results for men are given in the top panel and results for women are given in the bottom panel. For both men and women, the simulated distribution is almost coincident with the actual 2000 distribution. We stress that in other contexts, this need not be so—in such a case, it would be appropriate to compare a counter-factual such as $f_h^{t\to t'}(y)$ to $f_{g,h}^{t\to t'}(y)$, rather than to compare $f_h^{t\to t'}(y)$ to $f(y|T=t')$ (currently common practice).

The selection explanation: Methodology

The counter-factuals we have considered, like the original work of DFL, abstract from problems with selection. This will be valid if selection into income recipiency does not change between the two survey years (since we are explaining log incomes). In some cases, this will be an appropriate simplification. However, given the many changes in the South African labor market from 1995 to 2000, ignoring selection could lead to serious errors. For example, as we saw in Table 1, women

are somewhat more engaged in the labor market in 2000 than in 1995. However, we have little sense of what those who entered between 1995 and 2000 would have earned in 1995, had they worked. Similarly, men became less attached to the labor market over this period, and we do not know what those who exited would have earned, had they continued working. Before proceeding, we would like to emphasize the extent to which any selection story must be complex—as noted above, several population groups have distinct patterns of selection changes but common patterns of income decline.

To assess the potential relevance of the selection story, we implement a bounding procedure proposed by Lee (2004).[36] Using this procedure, we may bound the true 1995-2000 income difference under conservative assumptions.[37] We begin by describing Lee bounds in the context of a comparison of average log incomes between the two survey years, with no controls for individual level covariates. We then extend the analysis to cells defined by age, race, and gender. [38] Suppose that the data are generated according to

$$\begin{aligned} Y_i = &\quad \mu + \pi T_i + v_i \\ D_i = &\ 1(\delta_0 + \delta_2 T_i + w_i > 0) \end{aligned} \tag{15}$$

where Y_i is the log of individual i's income, T_i is an indicator variable for the year from which the data are observed, μ is a constant, and v_i is an individual-level idiosyncratic shock to income. While v_i and T_i are assumed orthogonal in the overall population, selection may result in income shocks being correlated with T_i for those individuals reporting positive income. Income, Y_i, is observed only if $D_i = 1$, and $D_i = 1$ only if $\delta_0 + \delta_2 T_i + w_i > 0$ where w_i is an idiosyncratic shock to income recipiency. While w_i is assumed orthogonal to T_i, w_i may be correlated with v_i. Intuitively, the latter correlation allows shocks to the level of income to be correlated with shocks to income recipiency—a somewhat natural correlation.

The parameter of interest is π, or the mean difference in log incomes for persons who would have received income in both survey years. In our sample, real incomes, conditional on being positive, are about 40 percent lower in 2000 than in 1995. If there were no selection, -0.40 would estimate π well. The methodology below accounts for selection by placing bounds on π.

[36] Lee focuses on the consequences of attrition for identification in the randomized control trial. However, the mathematical ideas apply equally to the analysis of changes in incomes between two time periods.

[37] As will become clear, however, the critical assumption of Lee bounds is a monotonicity condition, and this is probably only satisfied on a cell-by-cell basis in our application.

[38] This may be thought of as the discrete covariate approach. DiNardo (2002) suggests an approach that would extend Lee bounds to continuous covariates.

Because of the orthogonality of v_i and T_i, π is identified by $E[Y_i|T_i = 1] - E[Y_i|T_i = 0]$. However, because Y_i is only observed if $D_i = 1$, this is not an estimable quantity. It is feasible to estimate $E[Y_i|T_i = 1, D_i = 1] - E[Y_i|T_i = 0, D_i = 1]$, however, this quantity may be quite different from π to the extent that v_i and w_i are correlated, and to the extent that $\delta_2 \neq 0$.

The key insight of Lee (2004) is that it is straightforward to place approximate bounds on π using the fraction of the observations from each survey year for which we have a valid measure of income.[39] To see why this is so, suppose that $\delta_2 > 0$, or that the frequency with which income is observed rises from 1995 to 2000 (as is the case for women in our data).[40] According to the model in (15), average log income in the first survey year may be written

$$\begin{aligned} E[Y_i|T_i = 0, D_i = 1] &= \mu + E[v_i|T_i = 0, D_i = 1] \\ &= \mu + E[v_i|w_i > -\delta_0] \end{aligned}$$

Lee (2004) notes that we may write the average for the second year in an illuminating way:

$$\begin{aligned} E[Y_i|T_i = 1, D_i = 1] = & \quad \mu + \pi + E[v_i|T_i = 1, D_i = 1] \\ = & \quad \mu + \pi + E[v_i|w_i > -\delta_0 - \delta_2]\} \\ = & \quad p\{\mu + \pi + E[v_i| - \delta_0 - \delta_2 < w_i \leq -\delta_0]\} \\ & + (1-p)\{\mu + \pi + E[v_i|w_i > -\delta_0]\} \end{aligned}$$

where

$$p \equiv \frac{P(-\delta_0 - \delta_2 < w_i \leq -\delta_0)}{1 - P(w_i \leq -\delta_0 - \delta_2)}$$

That is, we can think of average log income in the second year as a weighted average of log income for the subpopulation that would have received income in both the first and second years, on the one hand, and for the subpopulation that would have received no income in the first year, but would have received income in the second year, on the other hand.

The idea behind this representation of the average for the second year is as follows. Suppose for the moment that we knew *which* individuals in the second year survey had $-\delta_0 - \delta_2 < w_i \leq -\delta_0$, or the subpopulation whose income recipiency is affected by the economic changes occurring between 1995 and 2000. Then we could delete them from the sample and compute average log income among

[39] This remains true under a heterogeneous treatment effects version of (15). The key aspect of (15) is not the particular functional form, but the monotonicity condition implied—namely that the "effect" of the survey year on the probability of being observed is of the same sign for all individuals. See Lee (2004) for details.

[40] The argument is analogous for the case that $\delta_2 < 0$. If $\delta_2 = 0$, then no selection correction is necessary.

the remaining individuals. That is, letting $M_i = 1$ indicate the event $-\delta_0 - \delta_2 < w_i \leq -\delta_0$, we have

$$E[Y_i|T_i = 1, D_i = 1, M_i = 0] = \mu + \pi + E[v_i|w_i > -\delta_0]$$

so that

$$E[Y_i|T_i = 1, D_i = 1, M_i = 0] - E[Y_i|T_i = 0, D_i = 1] = \pi$$

and the selection problem is circumvented.

In the absence of a good understanding of the selection process, it is not generally possible to identify the subpopulation with $M_i = 1$. However, consider for the moment the consequences of an extreme sort of selection, in which any given individual with $M_i = 1$ has a higher valued v_i than any given individual with $M_i = 0$. Then selecting on $M_i = 0$ is equivalent to trimming a fraction p of the observations with the highest v_i. Those with the highest v_i are equivalently those with the highest Y_i. This reasoning leads to the insight that the natural (but infeasible) estimator for $E[Y_i|T_i = 1, D_i = 1, M_i = 0]$, namely

$$\frac{\sum_{i=1}^{n} Y_i T_i D_i (1 - M_i)}{\sum_{i=1}^{n} T_i D_i (1 - M_i)},$$

can be no lower than

$$\frac{\sum_{i=1}^{[\![n(1-\hat{p})]\!]} Y_i T_i D_i}{\sum_{i=1}^{[\![n(1-\hat{p})]\!]} T_i D_i}$$

and can be no higher than

$$\frac{\sum_{i=[\![n\hat{p}]\!]}^{n} Y_i T_i D_i}{\sum_{i=[\![n\hat{p}]\!]}^{n} T_i D_i},$$

where the latter two sums (in order to minimize notational complexity) assume that the data are sorted in ascending order by Y_i, the operator $[\![\cdot]\!]$ rounds a real number to the nearest integer, and $\hat{p}$ is an estimate of p, namely

$$\hat{p} \equiv \frac{\hat{P}(D_i = 1|T_i = 1) - \hat{P}(D_i = 1|T_i = 0)}{\hat{P}(D_i = 1|T_i = 1)}$$

This simple trimming idea forms the building block of the bounds. Specifically, because the analysis above indicates that $E[Y_i|T_i = 1, D_i = 1, M_i = 0] - E[Y_i|T_i = 0, D_i = 1]$ is the parameter of interest, one estimates lower and upper bounds for π by

$$\widehat{\underline{\pi}} \equiv \frac{\sum_{i=1}^{[\![n(1-\hat{p})]\!]} Y_i T_i D_i}{\sum_{i=1}^{[\![n(1-\hat{p})]\!]} T_i D_i} - \frac{\sum_{i=1}^{n} Y_i (1 - T_i) D_i}{\sum_{i=1}^{n} (1 - T_i) D_i} \tag{16}$$

and

$$\widehat{\overline{\pi}} \equiv \frac{\sum_{i=[\![n\hat{p}]\!]}^{n} Y_i T_i D_i}{\sum_{i=[\![n\hat{p}]\!]}^{n} T_i D_i} - \frac{\sum_{i=1}^{n} Y_i (1 - T_i) D_i}{\sum_{i=1}^{n} (1 - T_i) D_i}, \tag{17}$$

respectively.[41] Appendix A provides two examples that illustrate how the trimming works in practice.

So far, we have discussed bounding when one observes a greater fraction of individuals in the second survey year (as is the case in our data for women). When one observes a greater fraction of individuals in the first survey year (as is the case in our data for men), it is then appropriate to trim observations from the first year.[42] Specifically, redefine the trimming fraction as

$$\hat{p} \equiv \frac{\hat{P}(D_i = 1|T_i = 0) - \hat{P}(D_i = 1|T_i = 1)}{\hat{P}(D_i = 1|T_i = 0)}$$

and note that now π is given by $E[Y_i|T_i = 1, D_i = 1] - E[Y_i|T_i = 0, D_i = 1, M_i = 0]$. Then one would form upper and lower bounds for π by

$$\widehat{\overline{\pi}} \equiv \frac{\sum_{i=1}^{n} Y_i T_i D_i}{\sum_{i=1}^{n} T_i D_i} - \frac{\sum_{i=1}^{\llbracket n(1-\hat{p}) \rrbracket} Y_i (1 - T_i) D_i}{\sum_{i=1}^{\llbracket n(1-\hat{p}) \rrbracket} (1 - T_i) D_i}$$

and

$$\widehat{\underline{\pi}} \equiv \frac{\sum_{i=1}^{n} Y_i T_i D_i}{\sum_{i=1}^{n} T_i D_i} - \frac{\sum_{i=\llbracket n\hat{p} \rrbracket}^{n} Y_i (1 - T_i) D_i}{\sum_{i=\llbracket n\hat{p} \rrbracket}^{n} (1 - T_i) D_i},$$

respectively.

The above description is formal, precise, and necessarily somewhat notationally complex. In contrast, the idea behind using trimming to place bounds on selection is actually pretty simple. Before adding the complexity of cell-by-cell or multidimensional trimming, it is perhaps helpful to provide an heuristic description of how the trimming approach to addressing selection applies to the case at hand.

Consider first the case of men. Results presented in section 3 showed that income recipiency declined for males overall from 1995 to 2000. To keep things simple, we'll pretend that the total sample sizes were the same in 1995 and 2000; it's just that fewer men reported positive incomes in 2000. Further, suppose that there were ten percent fewer men reporting positive incomes in 2000. Because we have fewer positive incomes in 2000, we want to know who had income in 1995 but did not have income in 2000. If we knew who these individuals were, we would exclude them from

[41] This formulation is distinct from that described in Lee (2004) in that it emphasizes the bounds as being satisfied for each sample size. (Lee's notation instead emphasizes the (correct) idea that the probability limit of the bounds will bound the probability limit of the infeasible estimator.) The treatment we have given here stresses that the bounds have a highly literal finite sample justification. This aids in intuition, but also conveys the important idea that standard errors do not need to be adjusted to accommodate sampling error in the trimming fraction $\hat{p}$ (this may be shown following the discussion in Section 6 of Newey and McFadden (1994)).

[42] This case corresponds to $\delta_2 < 0$.

the 1995 data and then be able to compare 1995 (after their exclusion) to 2000 and the selection problem would have disappeared. The bounds approach assumes first that the ten percent who no longer earned incomes were at the very top of the income distribution in 1995, so we trim from the top of the 1995 distribution. Having trimmed from the top, the mean of the resulting 1995 synthetic distribution is lower. If it was, say, 40 percent lower, the selection explanation (which motivated the trimming) could explain a 40 percent decline in incomes from 1995 to 2000. At the other extreme, it could be that the ten percent who earned incomes in 1995 but not in 2000 were at the bottom of the 1995 distribution. Trimming from the bottom of the 1995 distribution will result in a synthetic 1995 distribution with a higher mean. Since we know that men's mean incomes fell from 1995 to 2000, trimming from the bottom will only exacerbate the decline. In this case, selection will not explain the observed decline. By trimming from the top and then from the bottom, we place bounds on the role of selection. Intuitively, for selection to explain the decline in men's real incomes, it would have to be the case that the top ten percent of male earners in 1995 earned very high incomes.

Next consider the case for women. For women, income recipiency rose from 1995 to 2000. We'll again abstract from the different sample sizes and for purposes of this example, assume female income recipiency rose ten percent from 1995 to 2000. In this case, we ask, "Who are these women who earned incomes in 2000 but not in 1995?" At one extreme, they might comprise the top ten percent of the 2000 actual distribution. Trimming, then, will result in a synthetic distribution with a lower mean. If actual incomes fell by 40 percent, selection of this sort would exacerbate the decline, not make it disappear. Hence selection of this sort cannot explain the decline. At the other extreme, suppose those who earned income in 2000 but not in 1995 were at the very bottom of the 2000 distribution. Trimming these women will result in a synthetic 2000 distribution with a higher mean, hence the decline from 1995 to 2000 lessens. Selection, in this instance, could in theory explain the observed decline in women's mean log income.

Before examining results, recall that if selection is to explain the observed declines in income, it is a selection story that works very differently for men than for women. To fix ideas and stereotype, if selection explains the declines in incomes, it might be rich doctors who retire (but don't report any pension or other income) and poor uneducated women who start earning incomes.

The above analysis and discussion abstracts from covariates. Lee describes the trimming idea assuming that the observables have discrete support, in which case the above analysis may be conducted "cell-by-cell." This is essentially the approach we adopt. In interpreting our results, it is important to we note that in Lee's discussion, the focus was on the randomized control trial.

There, it was emphasized that the treatment being studied should never make some individuals more likely to be observed and other individuals less likely to be observed.[43] In our application, monotonicity is probably not satisfied globally, but may be satisfied on a cell-by-cell basis.[44]

This should be kept in mind when comparing our bounding results for broad subpopulations with those from more narrowly defined subpopulations. We turn now to results.

The selection explanation: Results

The question being asked is whether selection alone might explain the observed declines in individual real incomes. The descriptive analysis in Section 3 clearly indicates that there were changes in the likelihood of earning positive incomes between 1995 and 2000. Table 7 reports the results of our non-parametric analysis of the selection explanation.

Because Lee bounds are not yet common in the literature, we take care to describe Table 7 in some detail. The table reports selection-corrected log income differences. Following our convention of reporting results for men and women separately, we begin with the first two lines of Table 7. The first line reports results for men. In 1995, 68 percent of men reported positive incomes while in 2000 that figure fell to 61 percent. We trim, then, from 1995 since 1995 has the higher fraction of positive incomes. Suppose the men who left the ranks of those reporting positive income were at the very top of the 1995 distribution. Absent these men, the resulting 1995 trimmed distribution would have a mean log income of 9.73 and this is the figure reported in Column C. If the men who left between 1995 and 2000 had the lowest incomes, the mean of the trimmed distribution of log incomes is 10.19 and that is the figure reported in column E, labeled "upper bound." Selection in the case of men, then, could result in a mean log income between 9.73 and 10.19 with the actual 1995 figure being 9.96. In fact, mean log income fell in 2000 to 9.58 for men—the figure reported in column G. The actual difference in mean log incomes is -0.38 as reported in column J. Columns I and K report the bounds on this difference due to selection. Hence, mean log incomes could have been 0.60 lower (9.58 - 10.19) or might have only been 0.15 lower (9.58 - 9.73). Those bounds on the difference, importantly, do not bracket zero. Put another way, even the most extreme sort of selection cannot explain the observed decline in mean log incomes for men. The last column of

[43] In that case, there are additional complexities with combining the information on the bounds from all possible cells.

[44] For example, a literal interpretation of Lee would hold that the differential patterns for men and women rejects monotonicity.

Table 7 lists the fraction of the 1995 population who are men (0.48). Although Table 7 reports means, it should be noted that one could also report medians or any other quantile result.[45,46]

Next, consider the second line of Table 7. Here we report the results for women. As noted in columns A and B, the fraction reporting positive incomes increased so, in this case, we trim from the 2000 distribution to obtain the synthetic distributions from which we then compute bounds. Women's actual mean log income in 1995 was 9.44 (column D). That figure fell in 2000 to 9.08. To assess the scenario in which it was women at the top of the 2000 distribution who were missing in 1995, we trim these high income women in 2000. The mean log income of the synthetic distribution falls to 8.88 (column F). Similarly, to assess the scenario in which it was low income women who had positive incomes in 2000, we trim them. The resulting mean log income rises to 9.28 (column G). The first case exacerbates the difference and the difference in mean log incomes becomes -0.56. In the second case, selection mutes the difference, and the difference becomes -0.16 (instead of the actual difference of -0.37). Either way, selection alone cannot explain the decline in women's mean log income.

The next four rows investigate the selection explanation by race. For Blacks, the fraction earning positive incomes hardly changes so there is but a very limited role for selection and indeed the selection explanation is not very important. This finding, though, conflates the very different experiences of Black men and Black women. Findings for Coloureds are quite similar to those for Blacks. With men and women combined, there is not much of a role for the selection explanation. For Whites, on the other hand, the fraction earning positive incomes fell more substantially and selection could explain the rather modest 8 percent decline in mean log income. Taking into account the possible effects of selection, this difference could have been zero. That is, zero (no difference across years in mean log incomes) lies between the bounds reported in columns I and K. One should note, though, that this finding conflates the experiences of men and women and it applies to only 16 percent of the population.

In principle, one could compute selection-corrected log income differences at much finer granularity than just race *or* gender. In Appendix Table 4, we do so for age, race *and* gender cells; the structure of the table mimics that of Table 7. Appendix Table 4 contains three main messages.

[45] We report means because reporting only one summary statistic keeps the table more-or-less interpretable and we find that means work well given we are working with log incomes. Were we working with, say, income instead of log income, medians might be a better choice.

[46] However, the trimming rule depends on the choice of estimator. Suppose that one were interested in the difference in medians in the randomized control trial where a higher fraction of the control group was observed. Then one would compare the median for the treatment group to $\hat{Q}(0.5-\hat{p})$ and $\hat{Q}(0.5+\hat{p})$ where $\hat{Q}(\cdot)$ is the quantile function for the control group.

First, the declines in mean income are especially severe for younger cohorts. Young Black men and women comprise the two largest cohorts in Appendix Table 4 and these groups have two of the very steepest declines (0.52 for men and 0.61 for women 18-25 years old). As a general pattern, observed mean log income declines almost monotonically with age cohort within race and gender cells. These patterns were foreshadowed earlier by the age profiles of Figure 2. Second, for Blacks who comprise most of the population, the selection explanation just cannot explain the observed income declines, and this is true for both genders and all age cohorts. Third, there are specific cells for which selection might be important. On the other hand, many of these comprise miniscule fractions of the 1995 population. The youngest cohort of Black women is an exception. This cohort comprises a full 10 percent of the population and the upper bound suggests that selection could be explaining much (but still not all) of the huge (0.61) decline in log incomes. In this instance, the increase in young Black women entering the labor market really might have the very low incomes attributed to them in the trimming algorithm. Overall, though, while we acknowledge that selection may be of such a magnitude as to make it difficult to compare mean log incomes over time, we view it as unlikely that the observed declines in log incomes in South Africa are driven entirely by selection.

The finding that selection into income recipiency may be important in explaining income declines is corroborated by others. Casale and Posel (2002) document a large increase in Black female labor force participation in the post-apartheid era. They note that many of these Black females have battled to find employment and, as zero earners, are not reflected in either the distribution of log incomes nor are they reflected in the returns to education data. However, this increased supply of labor presumably placed downward pressure on wages in the occupations that Black females traditionally fill. (In this way, the selection explanation interacts with the returns explanation.) Also, Casale, Muller, and Posel (2004) note that many of these female labor market participants have been forced to work in more informal and less well compensated sections of the labor market. To the extent that Black female workers have begun to compete for the less skilled employment with Black male workers, the increased supply of Black female workers would exert downward pressure on male earnings (again illustrating a potential interaction between the returns and selection explanations).

5. Truths and Reconciliation

Real individual incomes in South Africa declined substantially from 1995 to 2000, and these declines are apparent throughout the distribution of incomes (see Figure 1 and Table 1). In general terms,

we have shown that changes in the distribution of endowments cannot explain the shift (see Figure 4), but that changes in returns to these endowments (see Figure 5) as well as selection into income recipiency (see Table 7) can in fact explain much of the decline. Underlying these very general explanations, several more specific findings emerge. First, the young (from ages 18 to the early 30's) were disproportionately impacted, and this is true for both men and women (see Figure 2). Second, the returns to education fell dramatically for Black men and women aged 18-25. More broadly, the return to education tended to fall for Blacks and rise or stay constant for non-Blacks. Young Whites in particular saw pronounced increases in the return to education. Fewer Whites, though, received income. Third, and more generally, we have shown that while selection into income recipiency may be empirically important for explaining the shift in incomes, it is selection of a complex variety. It is not a one-size-fits-all explanation as the experiences of men and women were quite different.

In this section, we step back and ponder what economic phenomena might underlie our findings. Our methodology cannot deliver the definitive answer. (It is not clear that any can.) We can, though, identify several factors that are at least consistent with our empirical findings as well as other factors that are not. We begin with *ex ante* plausible stories that are simply inconsistent with our findings.

As noted above, changes in endowments do not explain the decline. Two such changes are emigration (mostly of Whites) and changes in the rate of school leaving. While anecdotes of the South African doctor now in the U.K., Canada, or Australia are common, our results indicate that, rare or common, changes in the racial make-up of South African income recipients does not explain the shift in incomes. Others have noted that young Blacks are now obtaining higher levels of education, although those Blacks that do leave school are doing so before completing secondary education (Lam and Leibbrandt (2004)). These changing patterns of educational attainment, combined with strong convexities in the returns to education in South Africa through the 1990's (see Lam and Leibbrandt (2004) and Keswell and Poswell (2002)) could in principle lead to shifts in the distribution of individual incomes. In practice, they did not.

Another explanation that might explain the decline in incomes is that the data used are lousy. That is, the decline is simply an artifact and is not real. The "bad data" explanation is not a plausible explanation. The data are remarkably consistent in those dimensions where stability across time is expected, the survey instruments are nearly identical, the same agency collected both waves of the data, and the decline in incomes is corroborated by an observed increase in the share of household expenditure devoted to food.

It could also be that the real decline is explained quite simply by a story of sticky wages and rising prices.[47] We view the tremendous age and race heterogeneity in the decline as strong evidence against this story. That is, wages do not appear to be particularly sticky for Whites and the non-young.

There are several factors that are consistent with our findings. First, the 1995-2000 period was characterized by low employment growth and even declines in formal employment. There was substantial unemployment at the outset in 1995 and then net job creation on the order of 1.5 million to 2 million jobs (depending on the definition used) combined with an increase in job seekers on the order of 5 million to 6 million (Casale et al. (2004)). Simple partial equilibrium arguments suggest a downward pressure on wages. The decline in incomes, though, was not uniform. In general, Blacks fared worse than Whites, and this was true for both males and females, leading to a widening of the Black-White income gap. This is clearly illustrated in Appendix Figure 2. In that figure, the kernel density of the distribution of Black and White log incomes are graphed. It is clear that even in absolute terms, Blacks suffered larger declines. Since Black incomes tended to be smaller to begin with, the declines relative to incomes for Blacks are even greater.[48] The basic facts of unequal income declines seem to be that racial gaps widened, that the young fared worse than their elders, and that less educated workers took bigger hits than their better educated colleagues.

Undergirding these changing returns are changing labor demand patterns in the South African economy. Research on changing aggregate, sectoral, and occupational employment trends has highlighted the impacts of trade and technology in driving formal sector employment changes (see Bhorat and Hodge (1999) and Edwards (2000)). As in many developed countries, there is evidence of an increasingly skill-intensive and technology-intensive demand structure in the South African labor market. The same skill-biased technical change that is claimed to underlie increasing income inequality in some developed countries may also be occurring in South Africa.

In sum, then, the differentiation of returns between Blacks and Whites may well be predominantly driven by the fact that most Whites have education levels that benefited from any skill-biased technical change, whereas most Blacks do not. Any lingering difference in educational quality would exacerbate this. Coloureds are in a special position to benefit from these changes. Most Coloured students complete secondary education. Notably, Hoogeveen and Ozler (2004) find that Coloureds

[47] This is different than the bad data explanation which might argue the price deflators are somehow wrong.

[48] The bottom panel of Appendix Figure 2 shows real income increases for wealthier White women. There are very few of these women in the sample. More generally, the figure obscures the fact that there are about 7.5 times more Blacks than Whites.

were the only racial group to experience statistically significant declines in poverty between 1995 and 2000.

The findings that Blacks have fared worse in terms of income recipiency, that Black incomes fell more than White incomes, and that a change in returns (including those to race) explain most of the decline in income throughout the distribution might seem surprising for the five years *after* the new government took power. Any surprise, though, is muted by three factors. First, the longer-run process of de-racialization of the labor market had started in the mid-1970s and many of the easier gains may have already been embodied in the 1995 situation. For example, Hofmeyr (2000) shows that Black real wages grew notably faster than White real wages in the formal sector and in the non-primary sectors from 1975 to the beginning of the 1990's. Particularly important for our analysis, for less skilled Whites who owed their relatively privileged employment and earnings positions purely to the racial allocation of employment under apartheid, most of the downward adjustment in their earnings may have already taken place by 1995. Second, the low employment growth meant that new hires were opening up only very limited opportunities to reorganize the racial composition of the labor market. Third, the affirmative action policies that might have been expected to shore up Black earnings had not yet had much time, by 2000, to have an impact. It was during the 1994-1997 period that such policies were being put into place culminating in the 1998 Employment Equity Act (Maziya (2001)). On implementation, this act gave firms a *window* period to develop equity plans. Therefore, it may simply be unrealistic to expect affirmative action policies to have had much of an influence on the change in incomes, 1995-2000.

A combination of the slack labor market, skill biased technical change, and lingering discrepancies in the quality of schooling reconciles fairly well with our empirical findings. There are surely other competing explanations, but the above accords well with the data and with outside evidence. It is, in our view, a reasonable starting place for a discussion of why incomes have fallen.

6. Conclusion

This paper makes three substantive (as opposed to methodological) points. First, based on data from large national income and expenditure surveys conducted in 1995 and 2000, real individual incomes in South Africa declined from 1995 to 2000. This decline was substantial and occurred throughout the entire income distribution. Second, changes in the returns to individuals' endowments explain much of the decline. Third, selection into income recipiency changed from 1995 to 2000 and this too contributed to the observed decline in individual log incomes. The first point was illustrated using a careful merge of multiple data sets. The second and third points were documented by a non-parametric analysis that includes extensions to the approach of DiNardo et al.

(1996) and implementation of a trimming idea proposed by Lee (2004) for bounding the effect of selection.

The overall picture painted by the analysis is one of declining real individual incomes between 1995 and 2000. This is rather depressing, but it should not be all that surprising to those who read beyond government national accounts data. Several other researchers have found results that are broadly consistent with those reported here. Perhaps the best overview of these studies is the work of Fedderke, Manga, and Pirouz (2004) that plots measured household per capita income for each year from 1995 to 2000 as measured by six national household sample surveys. This study shows a marked decline in per capita incomes from 1995 to 1998. These incomes begin to rise again in 1999 and 2000 but they do not recover to 1995 levels.[49] Then, recent work by Ardington, Lam, Leibbrandt, and Welch (2005) on inequality and poverty using ten percent micro samples of the 1996 and 2001 population census reveals similarly bleak changes in household per capita incomes between 1996 and 2001. Since 2000, there have been releases of the Labour Force Survey (but no more recent income and expenditure surveys). It is reasonable to wonder if perhaps the more recent data paint a brighter picture. Preliminary analysis suggests that the answer is "no." Casale et al. (2004) analyze different measures of income than that used in this paper. They use data from 1995 to 2003, and they do not find substantial improvements in the 2000 to 2003 period. In summary, the empirical evidence from household surveys in South Africa generally seems to corroborate our basic findings.

We reviewed a number of plausible factors that could have led to this situation. Given the general reorganization of the South African economy after apartheid, it can be argued that the finding that real incomes did not rise does not surprise. The magnitude of the decline, on the other hand, is surprising, and, in our view, deserves further study. Whether the coming years will see South Africans' gains in political freedoms paired with improvements in economic well-being remains to be seen.

[49] See their Table 2 in particular.

Appendix A

Two Examples of the Trimming Rule

To develop intuition for the trimming rule, suppose that 100 individuals are sampled in each survey year, but that the income recipiency rate is different in the two survey years. This situation is depicted in the table below:

	Before Trimming		After Trimming	
	Year 0	Year 1	Year 0	Year 1
Observe Income	70	80	70	70
Do Not Observe Income	30	20	30	20
Total	100	100	100	90

For the "before trimming data", we see that $\hat{P}(D_i = 1|T_i = 0) = 0.7$, $\hat{P}(D_i = 1|T_i = 1) = 0.8$, and that the trimming fraction $\hat{p}$ is $(0.8 - 0.7)/0.8 = 0.125$. Consequently, we seek to trim 12.5% of the 80 observations with $D_i = T_i = 1$, or 10 observations. Before trimming, the total observation count (third row) was split evenly between the two years. However, the count of observations for those with income observed (first row) was not split evenly due to differential observation rates between the survey years. Trimming leads the selected sample of those with observed income to mimic the overall sample with respect to the equality of the number of observations coming from each survey year. This feature holds generally when sample sizes are different between the two surveys, but with respect to the fraction of the pooled observations that come from the two survey years. This situation is depicted in the table below:

	Before Trimming		After Trimming	
	Year 0	Year 1	Year 0	Year 1
Observe Income	70 (42%)	96 (58%)	70 (45%)	84 (55%)
Do Not Observe Income	30 (56%)	24 (44%)	30 (56%)	24 (44%)
Total	100 (45%)	120 (55%)	100 (48%)	108 (52%)

As before, $\hat{P}(D_i = 1|T_i = 0) = 0.7$ and $\hat{P}(D_i = 1|T_i = 1) = 0.8$, leading to a trimming fraction $\hat{p} = 0.125$. We therefore seek to trim 12.5% of the 96 observations with $D_i = T_i = 1$, or 12 observations. This leads to 84 post-trimming observations with $D_i = T_i = 1$. The trimming restores the 45-55 split in the pre-trimming overall sample (third row) to the post-trimming sample of those with observed income (first row).

References

Ardington, C., Lam, D., Leibbrandt, M., and Welch, M. (2005). The sensitivity of estimates of post-apartheid changes in South African poverty and inequality to key data imputations. CSSR Working Paper 05/106, University of Cape Town.

Autor, D. H., Katz, L. F., and Kearney, M. S. (2004). Trends in U.S. wage inequality: Re-assessing the revisionists. Unpublished Draft.

Bhorat, H., and Hodge, J. (1999). Decomposing shifts of labour demand in South Africa. *South African Journal of Economics*, *67*(3), 348–380.

Bourguignon, F., Ferreira, F., and Lustig, N. (2001). The microeconomics of income distribution dynamics: A comparative analysis of selected developing countries. Unpublished Draft.

Casale, D., Muller, C., and Posel, D. (2004). Two million net jobs: A reconsideration of the rise in employment in South Africa, 1995-2003. *South African Journal of Economics*, *72*(5), 978–1002.

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

Case, A., and Deaton, A. (1998). Large cash transfers to the elderly in South Africa. *Economic Journal*, *180*, 1330–1361.

Deaton, A. (2004). Measuring poverty in a growing world (or measuring growth in a poor world. Research Program in Development Studies, Woodrow Wilson School, Princeton University.

Deaton, A., and Paxson, C. (1998). Economies of scale, household size, and the demand for food. *Journal of Political Economy*, *106*(5), 897–930.

DiNardo, J. (2002). The human element and systematic bias in double-blind randomized trials with an application to chronic daily headache prophylaxis with tizanidine. Unpublished Draft.

DiNardo, J., Fortin, N., and Lemieux, T. (1996). Labor market institutions and the distribution of wages, 1973-1992: A semi-parametric approach. *Econometrica*, *64*(5), 1001–1044.

Edwards, L. (2000). Labour shedding output growth: Is trade the culprit?. *Trade and Industry Monitor*, *14*, 2–5.

Fedderke, J., Manga, J., and Pirouz, F. (2004). Challenging Cassandra: Household and per capita household income in the October Household Survey 1995-1999, income and expenditure surveys 1994 and 2000 and the Labour Force Survey 2000. Unpublished draft.

Gronau, R. (1974). Wage comparisons—a selectivity bias. *Journal of Political Economy*, *82*(6), 1119–1142.

Heckman, J. J. (1976). The common structure of statistical models of truncation, sample selection, and limited department variables and a simple estimator for such models. *Annals of Economic and Social Measurement*, *5*(4), 475–492.

Hofmeyr, J. (2000). The changing pattern of segmentation in the South African labour market. *Studies in Economics and Econometrics*, *24*(3), 109–128.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, *81*(396), 945–960.

Hoogeveen, J., and Ozler, B. (2004). Not separate, not equal: Poverty and inequality in post-apartheid South Africa. Unpublished Draft.

Johnston, J., and DiNardo, J. (1996). *Econometric Methods*. McGraw-Hill.

Juhn, C., Murphy, K. M., and Pierce, B. (1993). Wage inequality and the rise in returns to skill. *Journal of Political Economy*, *101*(3), 410–442.

Keswell, M. (2004). Education and racial inequality in post-apartheid South Africa. Working Paper No. 2004-02-008, Santa Fe Institute, Santa Fe, New Mexico.

Keswell, M., and Poswell, L. (2002). How important is education for getting ahead in South Africa. Working Paper No. 22, Centre for Social Science Research, University of Cape Town.

Lam, D., and Leibbrandt, M. (2004). What's happened to inequality in South Africa since the end of apartheid?. Unpublished Draft.

Lee, D. (2004). Trimming for bounds on treatment effects with missing outcomes. Unpublished Draft.

Lemieux, T. (2002). Decomposing changes in wage distributions: A unified approach. *Canadian Journal of Economics*, *35*(4), 646–688.

Levinsohn, J., and McCrary, J. (2005). Counterfactual income distributions. In Process.

Machado, J. A., and Mata, J. (2005). Counterfactual decomposition of changes in wage distributions using quantile regression. *Journal of Applied Econometrics*.

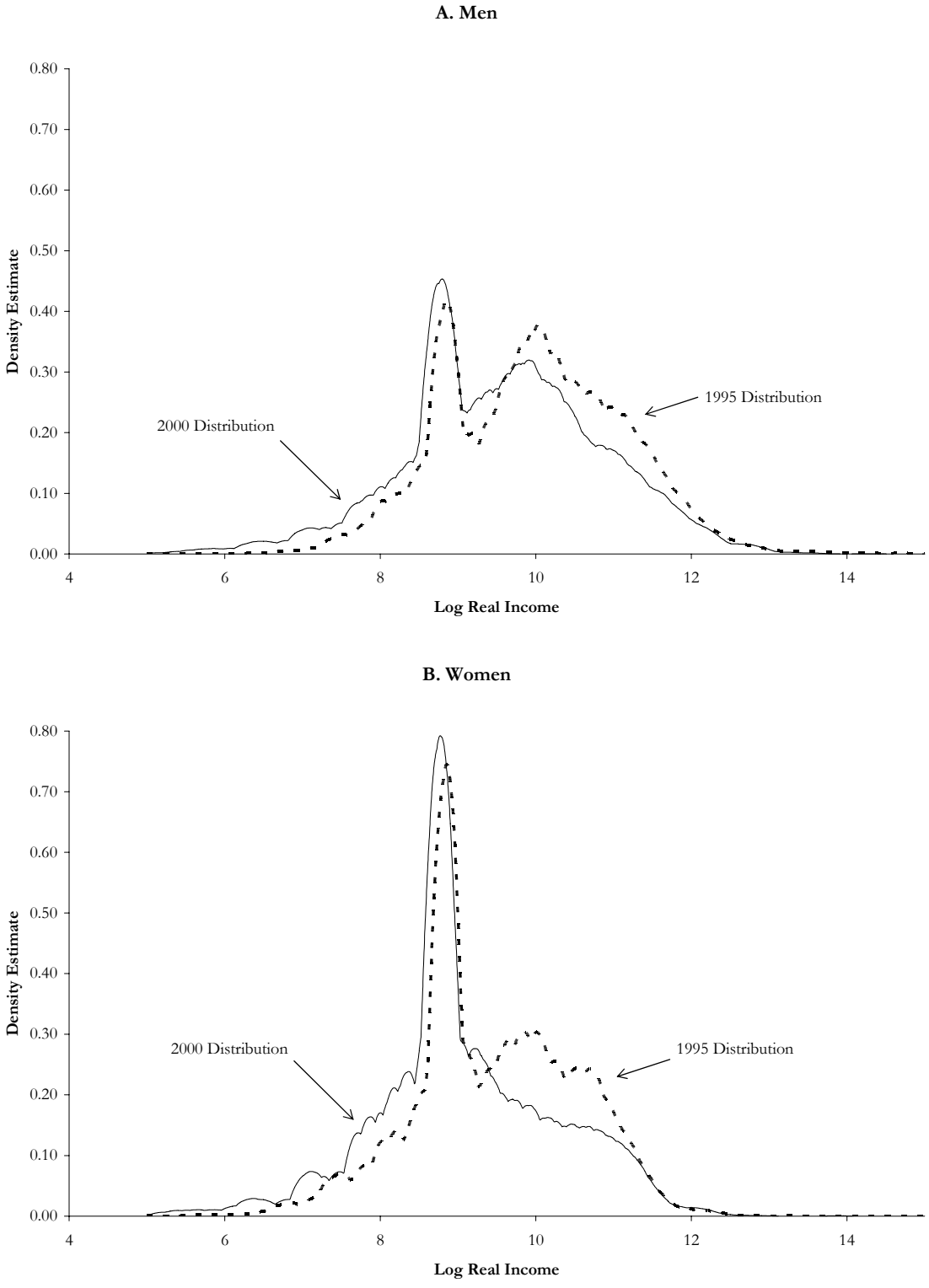
Maziya, M. (2001). Contemporary labour market poverty and policy in South Africa. In Bhorat, Leibbrandt, Maziya, der Berg, V., and Woolard (Eds.), *Fighting Poverty: Labour Markets and Inequality in South Africa*. UCT Press.

Newey, W. K., and McFadden, D. L. (1994). Large sample estimation and hypothesis testing. *Handbook of Econometrics*, 2111–2245.

UNDP (2003). *South Africa Human Development Report*. Oxford University Press, Cape Town.

Woolard, I. (2003). Impact of government programmes using adminisrative data sets: Social assistance grants. Unpublished Draft.

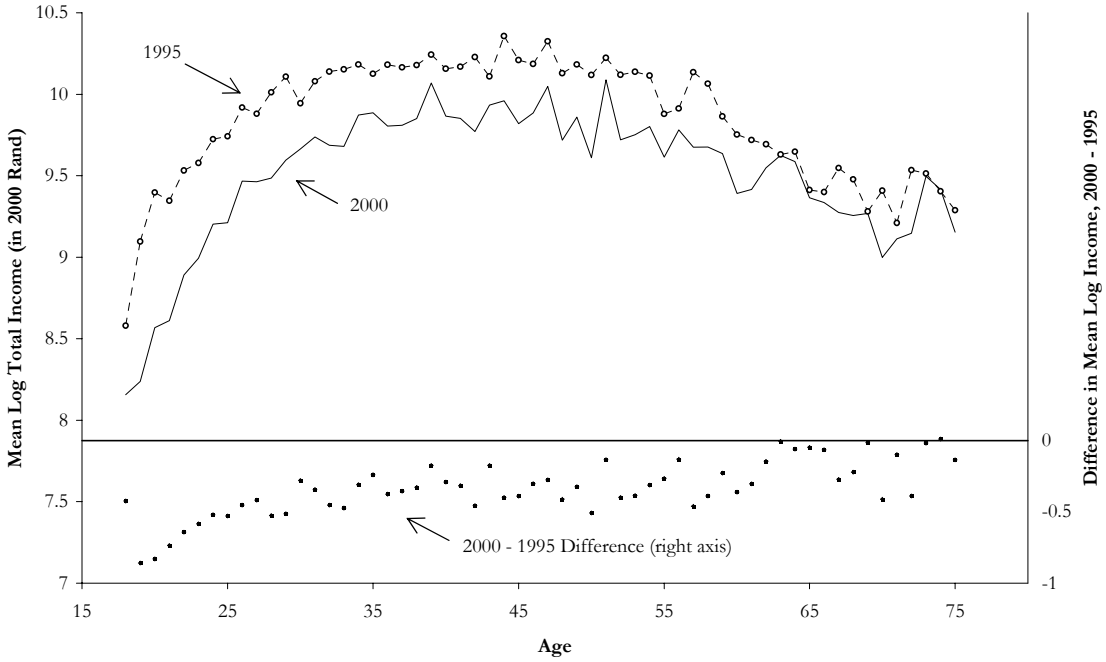
Figure 1. Distribution of Real Incomes in South Africa, 1995 and 2000



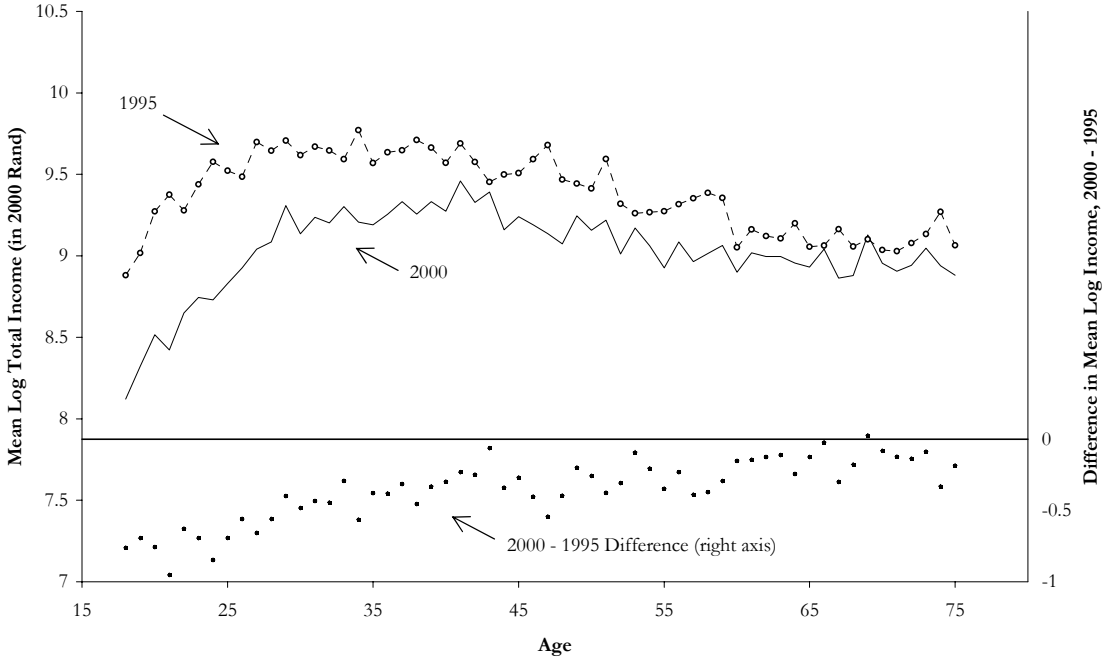
Note: Figure gives weighted kernel density estimates of log real total income (2000 Rand) for men (Panel A) and women (Panel B) in 1995 and 2000. All four density estimates use an Epanechnikov kernel and a bandwidth selector three-quarter the size of the Silverman (1986) rule-of-thumb (cf. Silverman's Equation 3.31). Sample sizes for 1995 and 2000 are 21,882 and 16,893 for men, respectively, and 18,868 and 17,776 for women, respectively.

Figure 2. Age Profiles of Mean Log Real Total Income, 1995 and 2000

A. Men



B. Women



Note: Figure gives average log real total income (2000 Rand) separately by year by gender for ages 18 to 75, inclusive. Sample sizes for the restricted age range for 1995 and 2000 are 21,247 and 16,387 for men, respectively, and 18,005 and 16,851 for women, respectively.

Figure 3. Food Share of Household Expenditures

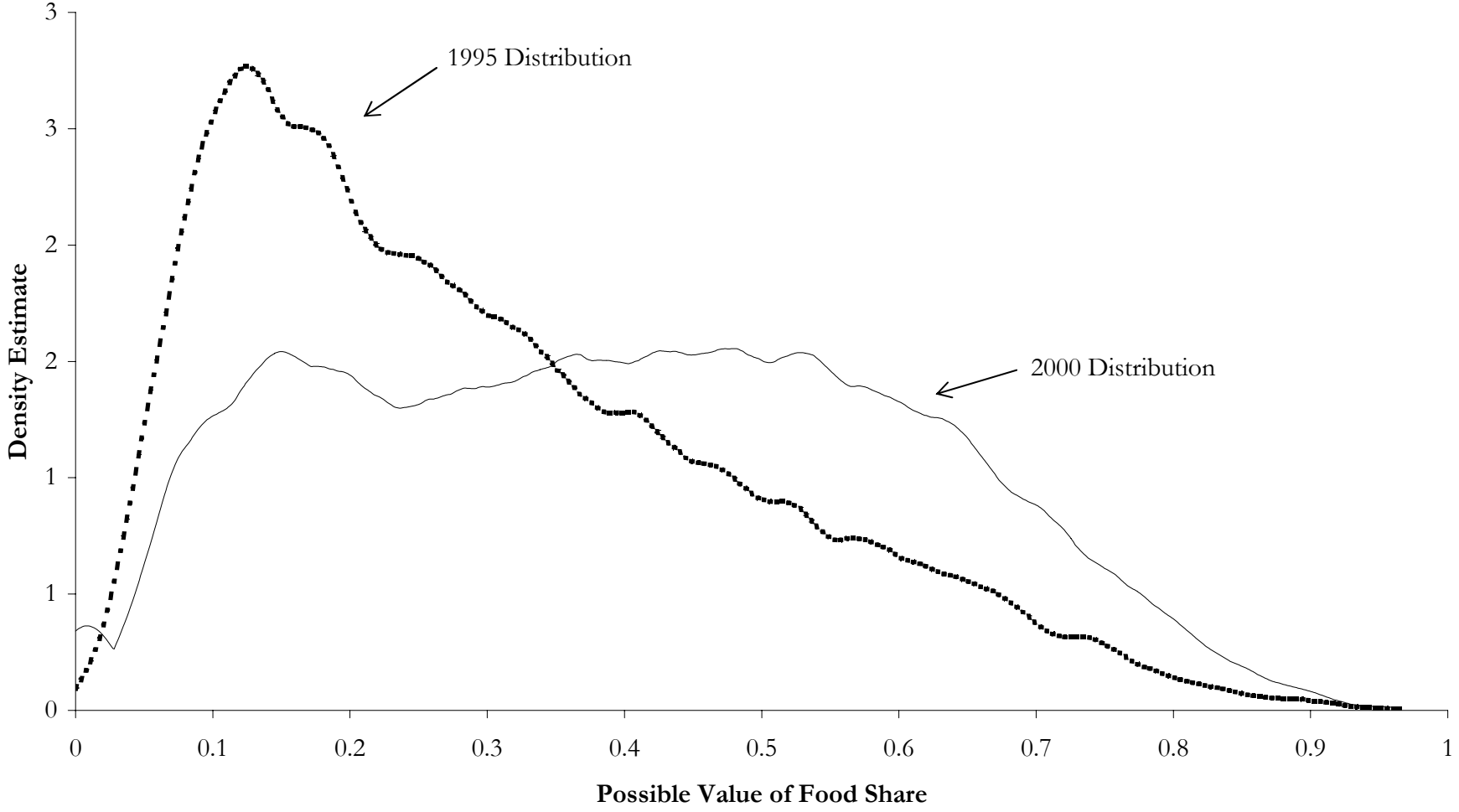
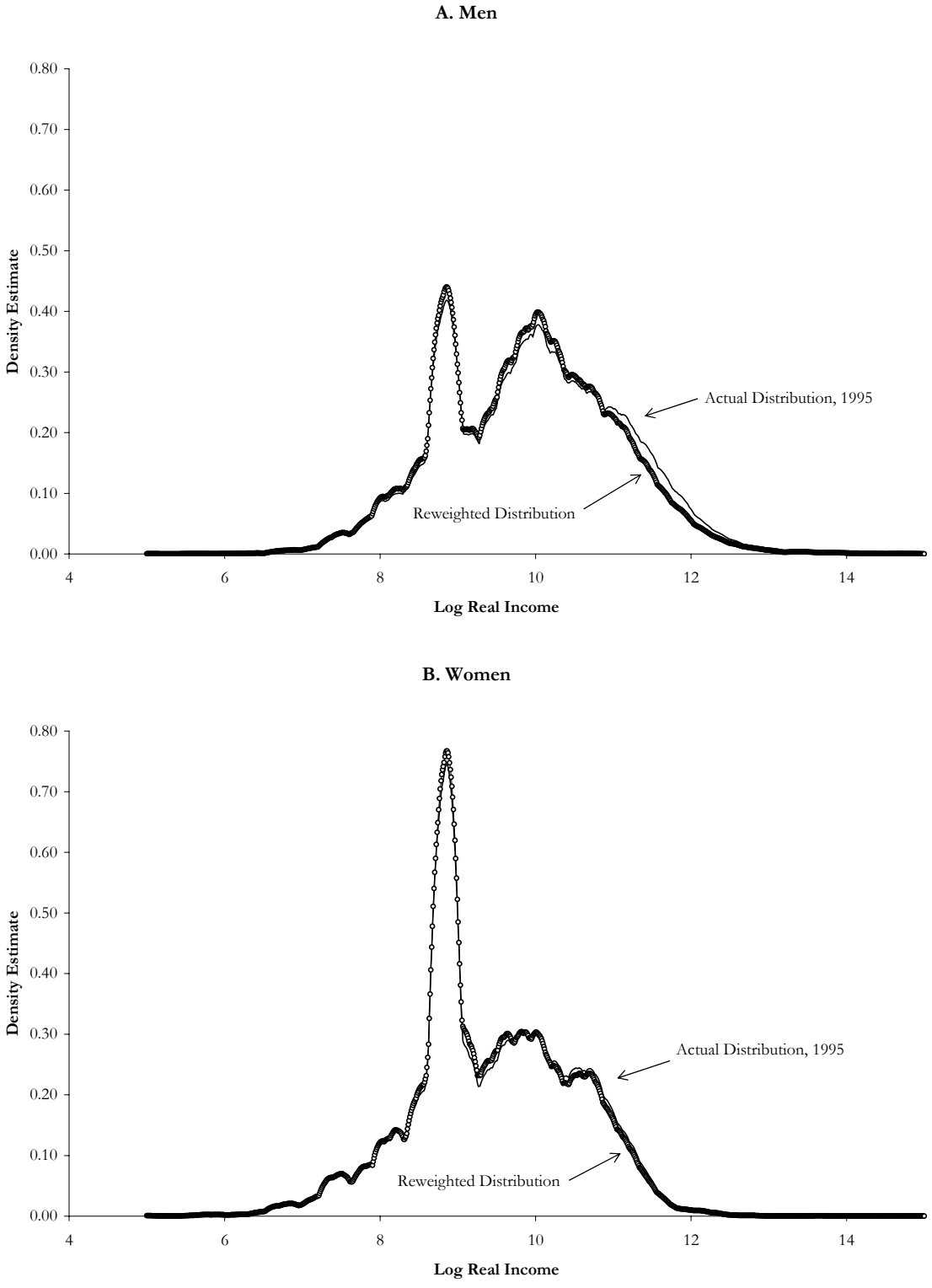
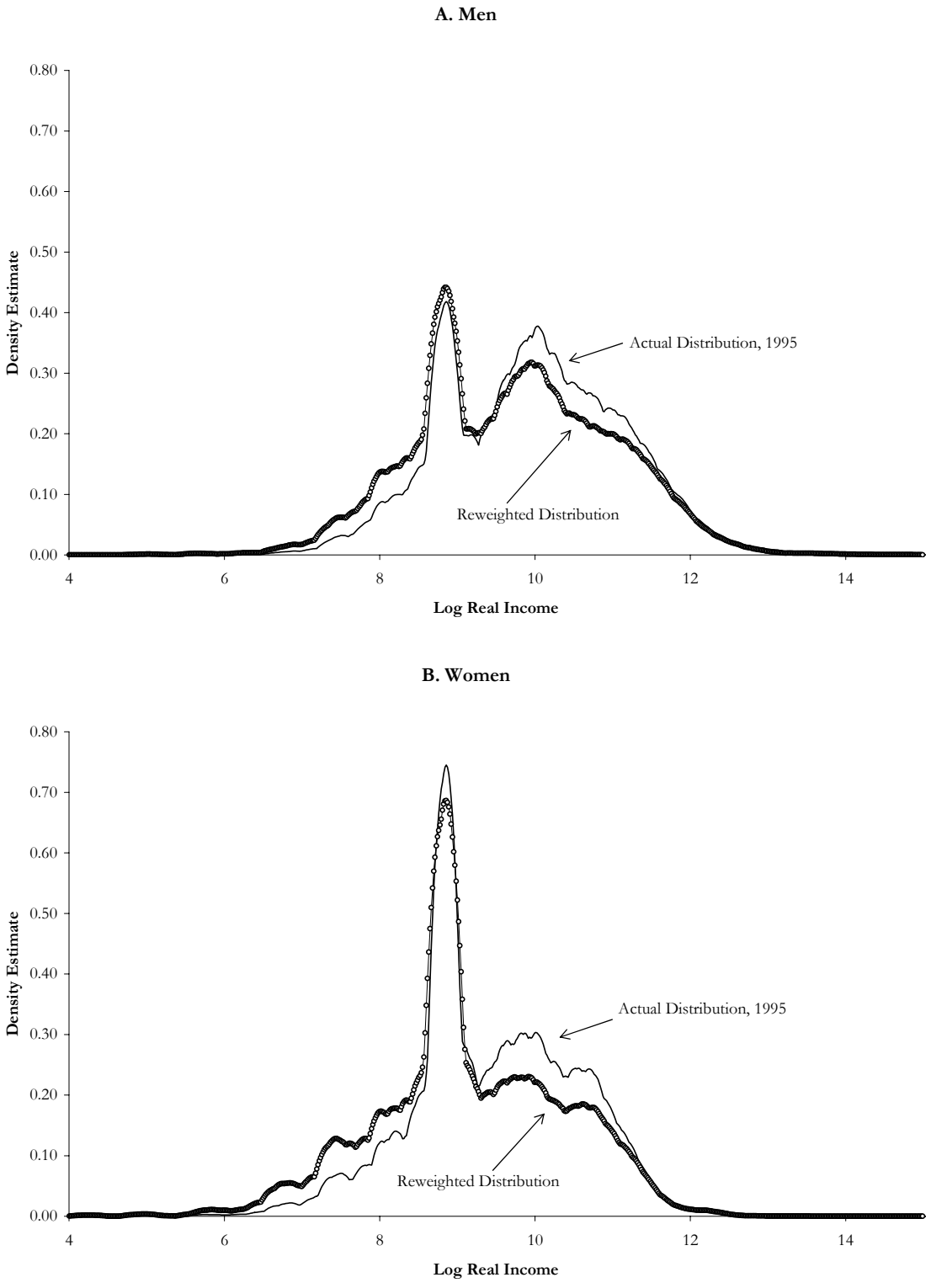


Figure 4. Omnibus Endowments Counterfactual, 1995 Data



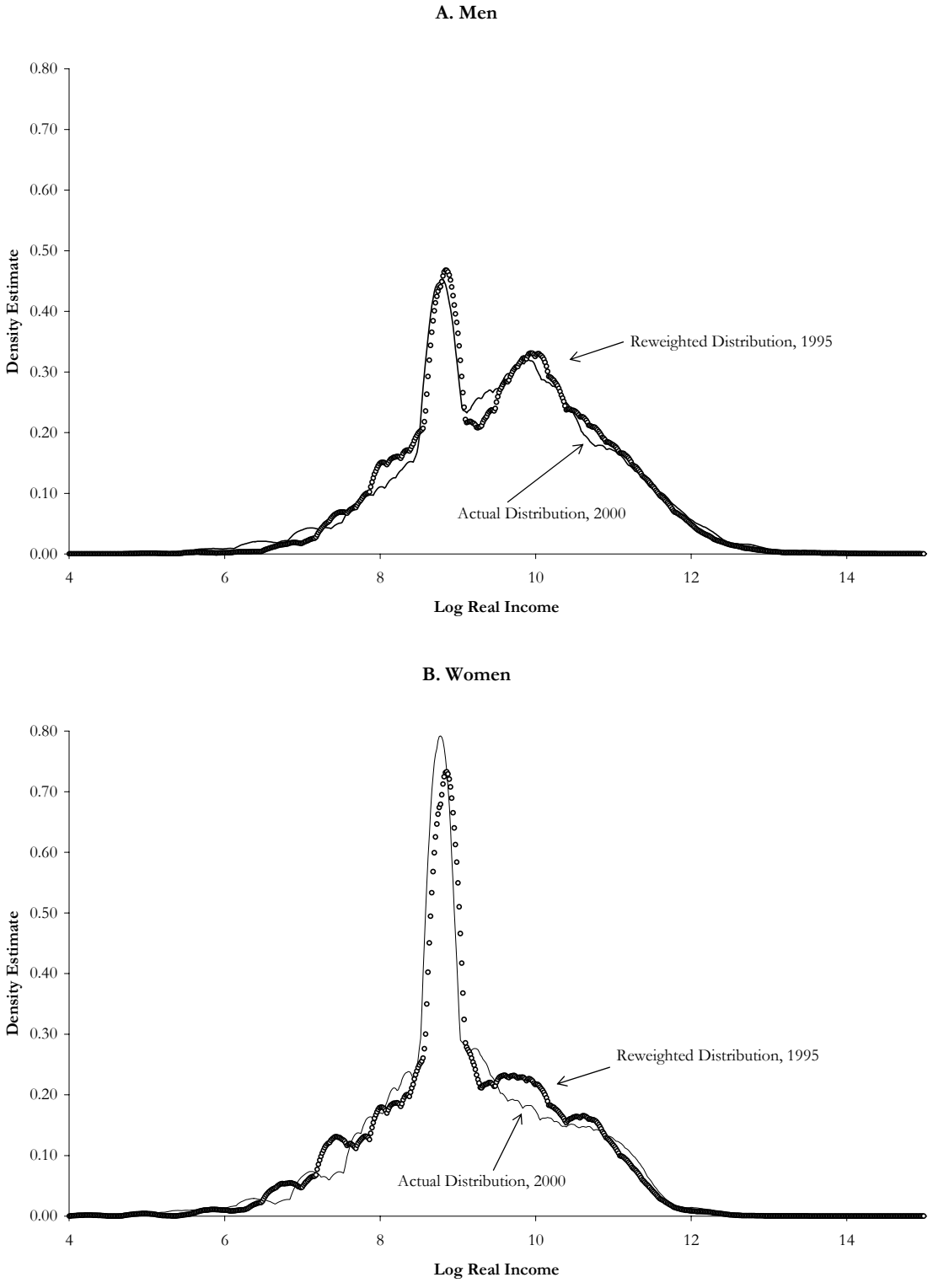
Note: Figure gives weighted kernel density estimates of log real total income (2000 Rand) for men (Panel A) and women (Panel B) in 1995, actual and reweighted to represent the counterfactual of changing endowments to the 2000 endowment distribution. Bandwidth rule and sample sizes are as described in the stub to Figure 1.

Figure 5. Omnibus Returns Counterfactual, 1995 Data



Note: Figure gives weighted kernel density estimates of log real total income (2000 Rand) for men (Panel A) and women (Panel B) in 1995, actual and reweighted to represent the counterfactual of changing returns to endowments to the 2000 returns mapping. Bandwidth rule and sample sizes are as described in the stub to Figure 1.

Figure 6. Comparability of Data Between 1995 and 2000



Note: Figure gives weighted kernel density estimates of log real total income (2000 Rand) for men (Panel A) and women (Panel B) in 2000 (actual) and a reweighted estimate for 1995 designed to represent the counterfactual of changing both 1995 endowments and 1995 returns to endowments to the 2000 endowments and the 2000 returns mapping. Bandwidth rule and sample sizes are as described in the stub to Figure 1.

Figure 7. Schematic View of Trimming Rule

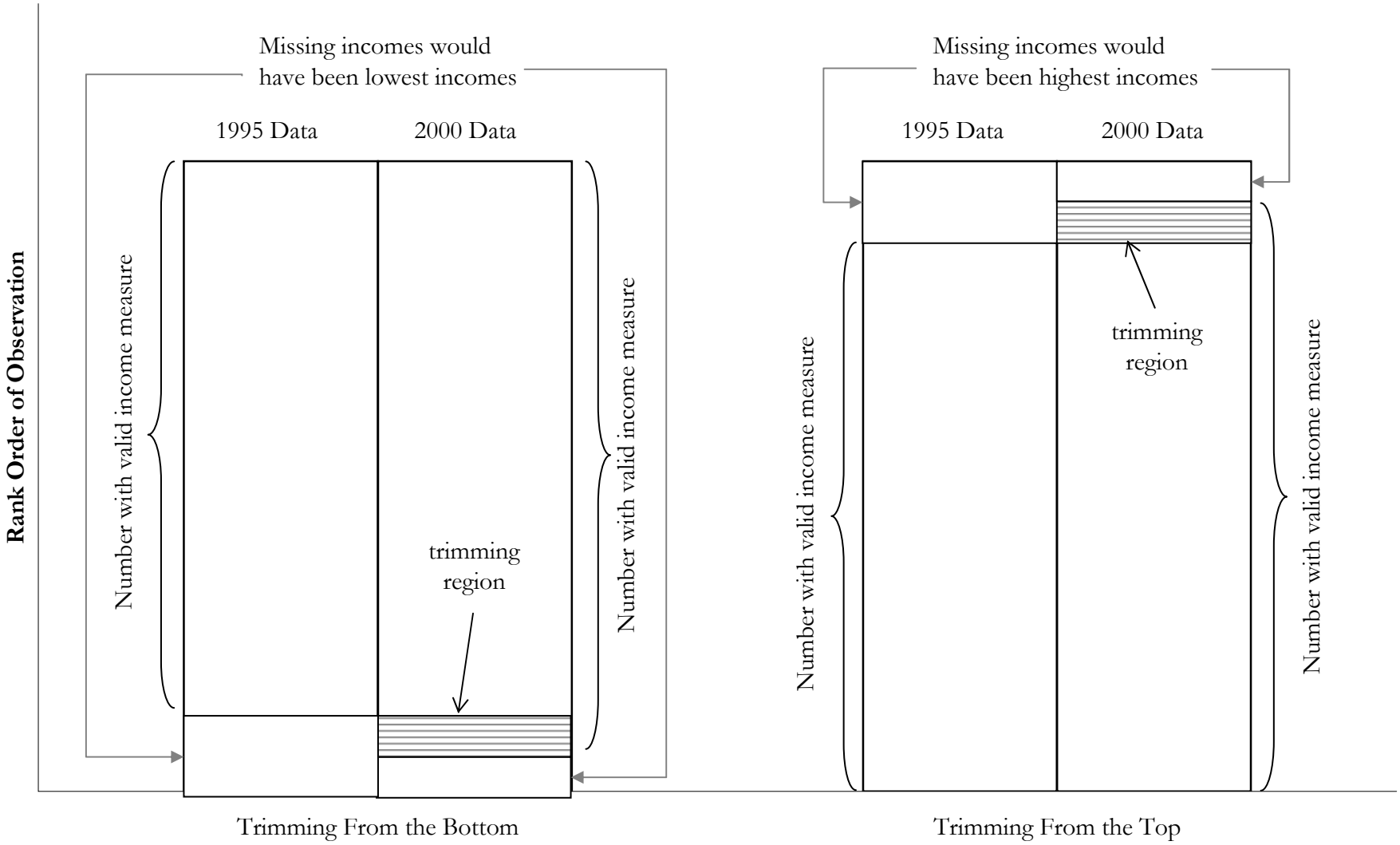


Table 1. Income Distributions in South Africa, 1995 and 2000 (in 2000 Rand)

		Men						Women					
		Total Income		Salary Income		Pension Income		All Income		Salary Income		Pension Income	
Row		1995	2000	1995	2000	1995	2000	1995	2000	1995	2000	1995	2000
		A	B	C	D	E	F	G	H	I	J	K	L
1	Fraction Positive	0.68	0.61	0.51	0.47	0.06	0.05	0.49	0.53	0.29	0.31	0.09	0.09
	Income Statistics												
2	Mean	42,378	32,286	39,106	37,696	7,711	8,111	21,742	17,561	26,535	24,801	7,107	6,849
3	S.D.	88,404	72,420	49,773	79,970	6,500	8,830	28,412	29,623	26,413	44,481	2,861	4,458
4	(90-10)/2.56	34,805	28,813	33,517	32,578	526	141	18,401	17,813	21,687	23,281	394	94
5	Skewness	14	27	8	22	12	6	7	11	3	13	12	12
6	Kurtosis	325	1,382	191	999	296	43	124	296	17	299	247	167
	Log Income Statistics												
7	Mean	9.96	9.58	10.04	9.78	8.86	8.83	9.44	9.08	9.67	9.34	8.83	8.76
8	S.D.	1.15	1.29	1.08	1.28	0.33	0.53	1.06	1.18	1.13	1.35	0.24	0.44
9	(90-10)/2.56	1.14	1.24	1.13	1.24	0.08	0.02	1.06	1.17	1.12	1.27	0.06	0.01
10	Skewness	0.08	-0.21	-0.27	-0.47	2.74	-2.53	-0.07	-0.10	-0.53	-0.52	0.98	-8.90
11	Kurtosis	3.06	3.81	3.09	4.49	27.13	57.38	3.16	3.94	2.81	4.30	30.92	146.61
	Income Percentiles												
12	1st	1,682	520	1,682	500	3,365	3,000	1,009	380	841	300	3,365	2,592
13	5th	3,365	1,800	3,365	2,100	5,047	6,120	2,243	1,200	2,019	1,200	5,215	6,000
14	10th	5,114	3,240	5,047	3,600	6,225	6,240	3,365	2,400	3,365	2,400	6,225	6,240
15	25th	8,412	6,480	11,777	8,400	6,730	6,480	6,730	4,800	7,066	4,800	6,730	6,480
16	50th	21,030	14,400	23,688	18,400	6,898	6,480	11,440	6,720	19,348	11,960	6,898	6,480
17	75th	47,780	33,600	49,210	40,600	6,898	6,480	27,339	18,000	38,056	31,920	6,898	6,480
18	90th	94,214	77,000	90,850	87,000	7,571	6,600	50,472	48,000	58,884	62,000	7,234	6,480
19	95th	134,592	120,000	126,180	132,000	11,070	12,000	68,978	68,400	75,708	82,488	8,378	6,840
20	99th	294,420	248,604	210,300	300,000	33,648	58,800	117,768	120,000	112,166	165,000	16,824	26,400
21	Observations	31,714	27,579	31,714	27,579	31,714	27,579	37,987	32,836	37,987	32,836	37,987	32,836

Note: Table gives real income statistics pertaining to income source specified in column headings for men and women in 1995 and 2000. South African prices increased by 39% from 1995 to 2000 (6.8% annual rate).

Table 2. Characteristics of the Population, 1995 and 2000

A. Men

Variable	Full Sample			Subsample with Income		
	1995	2000	Difference	1995	2000	Difference
	A	B	C	D	E	F
Age	37.1	36.6	-0.47	41.7	41.1	-0.63
	(15.1)	(15.0)	(0.14)	(14.5)	(14.9)	(0.17)
Education	8.4	8.5	0.03	8.3	8.4	0.04
	(4.2)	(4.1)	(0.04)	(4.4)	(4.3)	(0.05)
Urban	0.60	0.61	0.010	0.65	0.66	0.014
	(0.49)	(0.49)	(0.004)	(0.48)	(0.47)	(0.005)
Black	0.71	0.76	0.043	0.63	0.70	0.061
	(0.45)	(0.43)	(0.004)	(0.48)	(0.46)	(0.006)
White	0.16	0.12	-0.044	0.22	0.16	-0.065
	(0.37)	(0.32)	(0.004)	(0.41)	(0.36)	(0.005)
Coloured	0.10	0.09	-0.002	0.11	0.11	0.002
	(0.29)	(0.29)	(0.002)	(0.31)	(0.31)	(0.003)
Log Total Income				9.96	9.58	-0.38
				(1.15)	(1.29)	(0.01)
N	31,714	27,579	59,293	21,882	16,893	38,775

B. Women

Variable	Full Sample			Subsample with Income		
	1995	2000	Difference	1995	2000	Difference
	A	B	C	D	E	F
Age	38.4	38.2	-0.21	44.0	43.7	-0.32
	(16.0)	(16.0)	(0.13)	(16.3)	(16.5)	(0.19)
Education	7.9	8.1	0.27	7.7	7.7	0.03
	(4.4)	(4.3)	(0.04)	(4.7)	(4.6)	(0.06)
Urban	0.53	0.58	0.051	0.59	0.60	0.014
	(0.50)	(0.49)	(0.004)	(0.49)	(0.49)	(0.006)
Black	0.72	0.77	0.042	0.66	0.73	0.074
	(0.45)	(0.42)	(0.004)	(0.47)	(0.44)	(0.006)
White	0.15	0.11	-0.047	0.20	0.13	-0.073
	(0.36)	(0.31)	(0.003)	(0.40)	(0.33)	(0.005)
Coloured	0.09	0.10	0.002	0.11	0.11	-0.005
	(0.29)	(0.30)	(0.002)	(0.32)	(0.31)	(0.003)
Log Total Income				9.44	9.08	-0.37
				(1.06)	(1.18)	(0.01)
N	37,987	32,836	70,823	18,868	17,776	36,644

Notes: Columns A, B, D, and E give weighted means (standard deviations) of key variables for specified subpopulations. Columns C and F give weighted difference in means (White standard errors) for key variables. Full sample is those with non-missing age, sex, race, education, and statistical weight. Subsample is the subset of the full sample with positive individual income.

Table 3. Mincer Regressions for Total Income

	Men			Women		
	1995	**2000**	**Difference**	**1995**	**2000**	**Difference**
Variable						
White	0.979	1.176	0.198	0.602	0.984	0.382
	(0.019)	(0.034)	(0.046)	(0.022)	(0.036)	(0.051)
Indian	0.576	0.635	0.058	0.528	0.578	0.050
	(0.028)	(0.057)	(0.066)	(0.035)	(0.054)	(0.076)
Coloured	0.098	0.298	0.200	0.105	0.315	0.210
	(0.017)	(0.027)	(0.041)	(0.018)	(0.028)	(0.041)
Age	0.096	0.114	0.018	0.042	0.066	0.023
	(0.003)	(0.005)	(0.006)	(0.003)	(0.003)	(0.004)
Age-Squared/100	-0.097	-0.108	-0.012	-0.038	-0.050	-0.012
	(0.004)	(0.005)	(0.007)	(0.003)	(0.003)	(0.004)
Over Pension Age	-0.024	0.032	0.056	-0.075	-0.078	-0.003
	(0.041)	(0.057)	(0.069)	(0.029)	(0.033)	(0.046)
Education	0.118	0.112	-0.006	0.116	0.113	-0.004
	(0.002)	(0.003)	(0.004)	(0.002)	(0.002)	(0.004)
Observations	21,882	16,893	38,775	18,868	17,776	36,644

Note: Table gives coefficients for weighted least squares regressions of log annual total income on individual attributes separately for men and women for 1995 and 2000, weighted by the person weight. Standard error (parentheses) are so-called clustered standard errors at the level of the household. Columns labelled "difference" were obtained from a stacked regression using data from both survey years, taking care to normalize the person weights to sum to one separately for each survey year.

Table 4. Income Recipiency Rates

Group	Both Genders			Men			Women		
	1995	2000	Difference	1995	2000	Difference	1995	2000	Difference
All	0.58	0.57	-0.01 *	0.68	0.61	-0.07 *	0.49	0.53	0.04 *
Population Group									
Whites	0.77	0.71	-0.06 *	0.92	0.79	-0.13 *	0.64	0.63	0.00
Blacks	0.52	0.53	0.01 *	0.60	0.56	-0.05 *	0.45	0.51	0.06 *
Coloureds	0.68	0.66	-0.02 *	0.77	0.72	-0.05 *	0.59	0.60	0.01
Indians	0.64	0.62	-0.02	0.83	0.72	-0.11 *	0.44	0.52	0.07 *
Age									
18-25	0.26	0.26	0.01	0.29	0.29	0.00	0.22	0.24	0.02 *
26-30	0.55	0.54	-0.02	0.68	0.61	-0.07 *	0.44	0.47	0.03 *
31-35	0.66	0.64	-0.02 *	0.79	0.71	-0.08 *	0.53	0.58	0.04 *
36-45	0.71	0.69	-0.02 *	0.86	0.76	-0.10 *	0.57	0.63	0.06 *
46-60	0.73	0.69	-0.03 *	0.89	0.76	-0.13 *	0.59	0.64	0.05 *
over 60	0.86	0.87	0.01	0.93	0.86	-0.06 *	0.81	0.87	0.06 *
Education									
Less Than 11 Years	0.58	0.57	0.00	0.68	0.60	-0.08 *	0.49	0.55	0.06 *
11-12 Years	0.51	0.50	-0.02	0.61	0.56	-0.05 *	0.42	0.44	0.02
More Than 12 Years	0.81	0.75	-0.07 *	0.88	0.79	-0.09 *	0.74	0.71	-0.03 *

Note:

Table gives income recipiency rates for various subpopulations. Columns labelled "difference" give the coefficient on a dummy for 2000 in a pooled regression of data for both years using observations only from the specified cell. The asterisk denotes statistical significance at the 5% level, and is based on the t-ratio test using standard errors clustered at the household-year level.

Table 5. Marginal Effects for Logit Models for Income Recipiency

	Men			Women		
	1995	**2000**	**Difference**	**1995**	**2000**	**Difference**
Variable						
White	0.286	0.139	-0.148	0.076	0.014	-0.063
	(0.013)	(0.015)	(0.020)	(0.013)	(0.016)	(0.021)
Indian	0.189	0.105	-0.084	-0.045	-0.047	-0.003
	(0.016)	(0.024)	(0.029)	(0.018)	(0.018)	(0.026)
Coloured	0.152	0.133	-0.019	0.135	0.078	-0.057
	(0.010)	(0.013)	(0.016)	(0.010)	(0.011)	(0.015)
Age	0.055	0.048	-0.006	0.035	0.032	-0.003
	(0.001)	(0.001)	(0.002)	(0.001)	(0.001)	(0.002)
Age-Squared/100	-0.054	-0.047	0.007	-0.030	-0.025	0.005
	(0.002)	(0.002)	(0.003)	(0.001)	(0.001)	(0.002)
Over Pension Age	0.399	0.404	0.005	0.239	0.218	-0.021
	(0.042)	(0.042)	(0.059)	(0.019)	(0.023)	(0.030)
Education	0.001	0.005	0.003	0.010	0.007	-0.003
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Log-Likelihood	-14,065	-15,575		-23,219	-19,953	
Observations	31,714	27,579		37,987	32,836	

Note: Table gives marginal effects for coefficients from weighted logit models of total income recipiency separately for men and women for 1995 and 2000, weighted by the person weight. Standard errors (parentheses) are so-called clustered standard errors at the level of the household and were calculated by the so-called delta-method.

Table 6. Comparability of Data

		Men			Women		
Outcome	**Cohort**	**1995**	**2000**	**Difference**	**1995**	**2000**	**Difference**
White	1931-1935	0.26	0.25	0.00	0.20	0.14	-0.06 *
	1936-1940	0.23	0.21	-0.02	0.20	0.17	-0.03
	1941-1945	0.22	0.20	-0.02	0.20	0.16	-0.04 *
	1946-1950	0.21	0.17	-0.04	0.20	0.16	-0.03
	1951-1955	0.18	0.15	-0.03	0.16	0.14	-0.02
	1956-1960	0.16	0.13	-0.03	0.14	0.12	-0.02
	1961-1965	0.14	0.14	-0.01	0.13	0.11	-0.03 *
	1966-1970	0.13	0.09	-0.03 *	0.12	0.10	-0.02 *
	1971-1975	0.11	0.08	-0.03 *	0.11	0.07	-0.03 *
Indian	1931-1935	0.03	0.04	0.01	0.03	0.03	0.01
	1936-1940	0.03	0.03	0.00	0.03	0.03	0.00
	1941-1945	0.04	0.04	0.01	0.03	0.04	0.01
	1946-1950	0.03	0.05	0.01	0.03	0.04	0.01
	1951-1955	0.04	0.04	0.01	0.03	0.04	0.01
	1956-1960	0.03	0.04	0.01	0.03	0.03	0.01
	1961-1965	0.03	0.03	0.00	0.03	0.03	0.00
	1966-1970	0.03	0.03	0.00	0.03	0.03	0.01
	1971-1975	0.03	0.03	0.01	0.02	0.02	0.00
Coloured	1931-1935	0.08	0.10	0.02	0.08	0.09	0.00
	1936-1940	0.08	0.09	0.01	0.09	0.08	-0.01
	1941-1945	0.09	0.09	0.01	0.09	0.11	0.02
	1946-1950	0.09	0.11	0.02	0.09	0.10	0.01
	1951-1955	0.09	0.09	0.00	0.10	0.10	0.00
	1956-1960	0.10	0.10	0.00	0.10	0.11	0.01
	1961-1965	0.11	0.10	0.00	0.11	0.11	0.00
	1966-1970	0.11	0.10	-0.01	0.10	0.10	0.00
	1971-1975	0.10	0.09	-0.01	0.10	0.09	-0.01
Education	1931-1935	6.5	6.0	-0.6	5.1	4.6	-0.4
	1936-1940	6.5	6.1	-0.4	5.8	5.1	-0.7 *
	1941-1945	7.3	6.7	-0.6 *	6.1	5.9	-0.2
	1946-1950	7.7	6.9	-0.9 *	6.8	6.4	-0.4 *
	1951-1955	8.1	7.4	-0.7 *	7.3	6.7	-0.6 *
	1956-1960	8.8	8.1	-0.6 *	8.1	7.6	-0.5 *
	1961-1965	9.0	8.8	-0.2	8.7	8.3	-0.3 *
	1966-1970	9.4	9.1	-0.3 *	9.4	9.3	-0.1
	1971-1975	9.7	9.9	0.1	9.8	9.9	0.2

Note: Table gives cohort-specific means of specified variables for men and women in 1995 and 2000. Column labelled "difference" is the coefficient on a dummy for 2000 in a pooled regression of data for both years using observations only from the specified cohort and taking care to normalize the person weights to sum to one separately for each survey year. The asterisk denotes statistical significance at the 5% level, and is based on the t-ratio test using standard errors clustered at the household-year level.

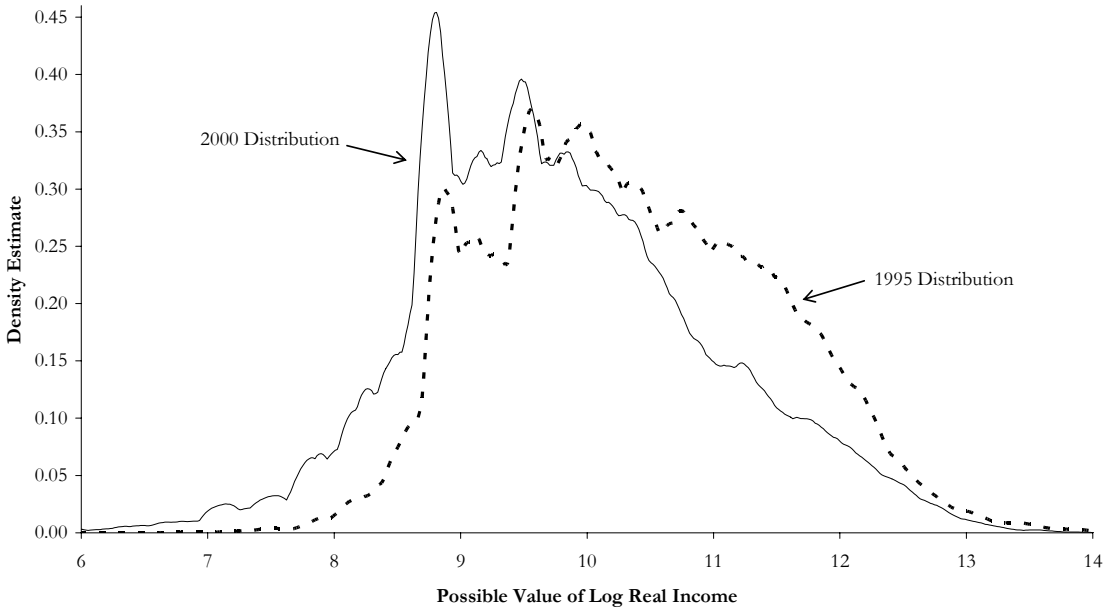
Table 7. Selection-Corrected Log Income Differences

Population Group	Fraction Positive Income		1995			2000			Difference			1995 Pop. Fraction
	1995	2000	Lower Bound	Actual	Upper Bound	Lower Bound	Actual	Upper Bound	Lower Bound	Actual	Upper Bound	
	A	B	C	D	E	F	G	H	I	J	K	L
Men	0.68	0.61	9.73	9.96	10.19		9.58		-0.60	-0.38	-0.15	0.48
Women	0.49	0.53		9.44		8.88	9.08	9.28	-0.56	-0.37	-0.16	0.52
Blacks	0.52	0.53		9.38		8.95	9.00	9.07	-0.43	-0.38	-0.31	0.72
Whites	0.77	0.71	10.67	10.82	11.01		10.74		-0.27	-0.08	0.07	0.16
Indians	0.68	0.66	9.48	9.55	9.62		9.43		-0.19	-0.12	-0.05	0.10
Coloureds	0.64	0.62	10.30	10.37	10.44		10.03		-0.41	-0.34	-0.26	0.03

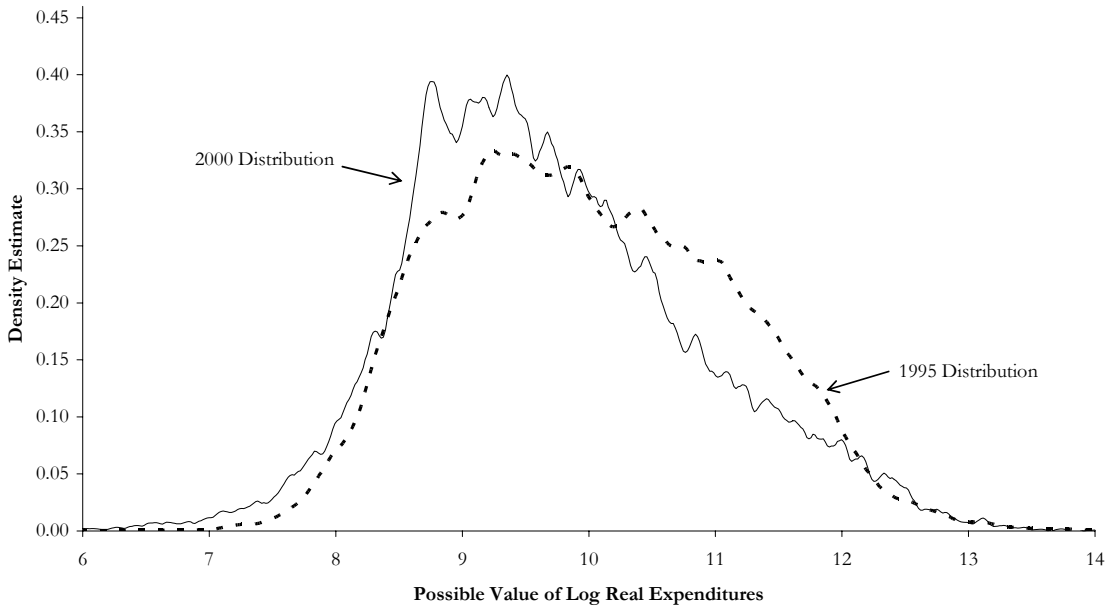
Note: Each entry in the table gives a statistic for a specific population group described in the first column. Columns A and B give the fraction of persons with positive income in 1995 and 2000, respectively. Columns C, D, and E pertain to average 1995 log incomes for those with positive incomes. Column D gives the actual average, while column C gives the lower bound (when applicable) for this average and column E gives the upper bound (when applicable). See text for description of bounds. When the entry is left blank, no trimming for 1995 is appropriate. Columns F, G, and H give analogous information for average 2000 log incomes. Columns I, J, and K are also organized analogously, but pertain to the 2000-1995 log difference. For context, column L gives the fraction of the overall 1995 population represented by the given cell.

Appendix Figure 1. Log Real Household Income and Expenditures

A. Log Household Real Income (2000 Rand)



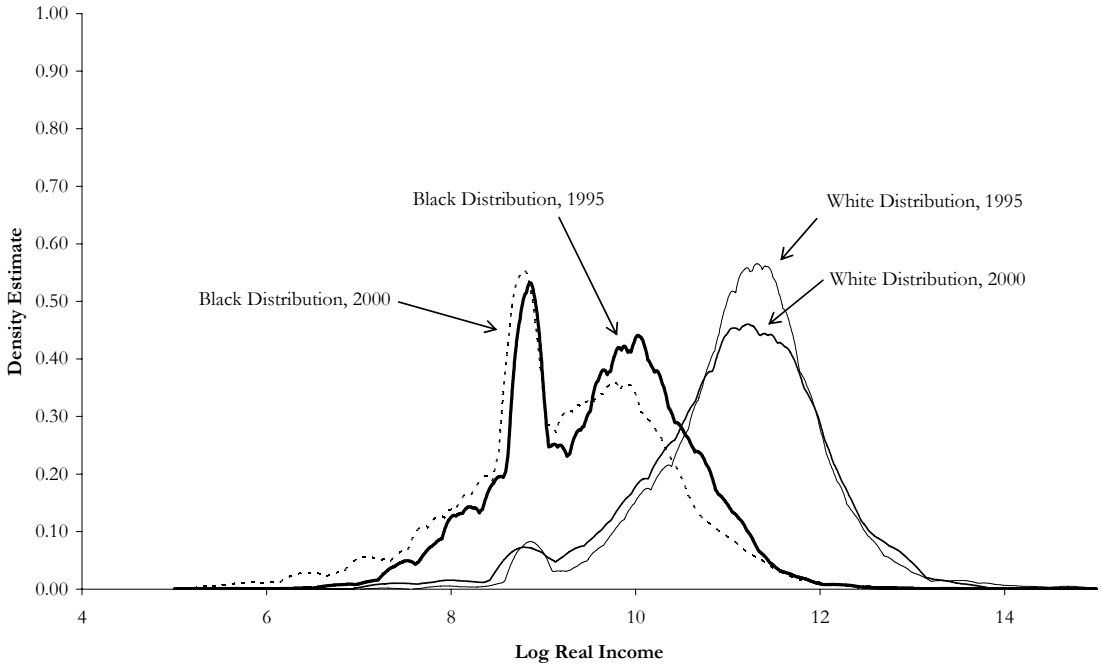
B. Log Household Real Expenditures (2000 Rand)



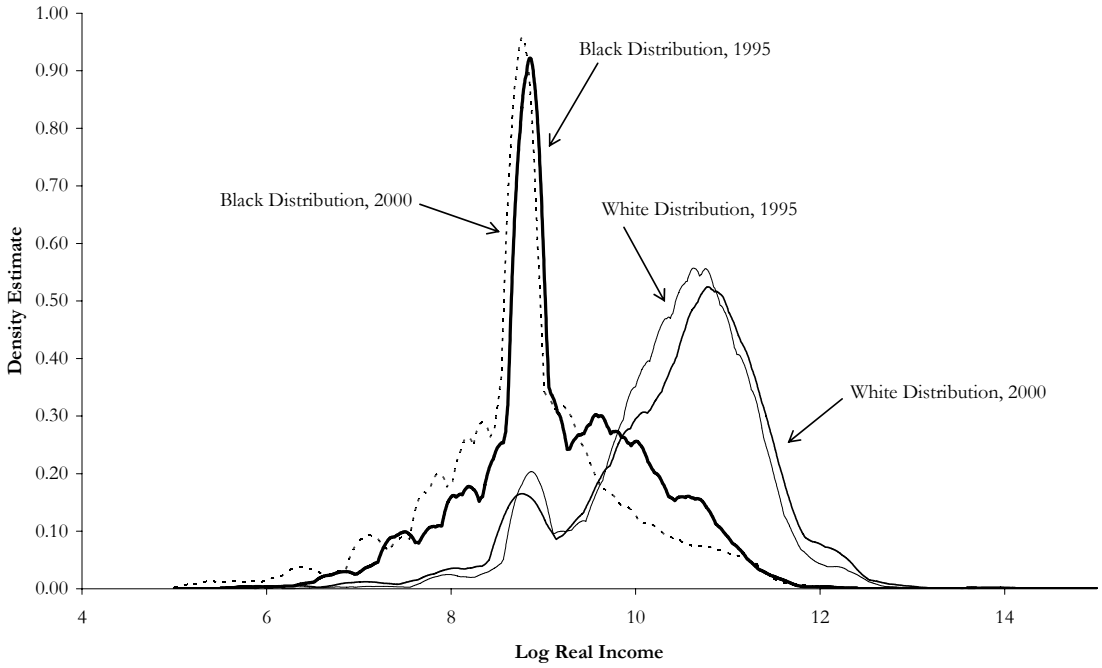
Note: Top panel of figure gives density estimates of log real household income (2000 Rand) separately by year. Bottom panel gives density estimates of log real household expenditures (2000 Rand) separately by year.

Appendix Figure 2. Log Real Income by Race by Year by Gender

A. Men



B. Women



Note: Top panel of figure gives density estimates of log real household income (2000 Rand) separately by race by year for men. Bottom panel is identical but is for women.

Appendix Table 1. Income Recipiency Rates by Age-Race-Sex-Education Cells

Population Group	Age	Less than 11 Years			11-12 Years Education			More than 12 Years		
		1995	2000	Change	1995	2000	Change	1995	2000	Change
Black Men	18-25	0.23	0.23	0.00	0.20	0.25	0.05 *	0.36	0.46	0.10
	26-30	0.60	0.53	-0.07 *	0.60	0.56	-0.04 *	0.76	0.66	-0.11 *
	31-35	0.69	0.65	-0.05 *	0.74	0.67	-0.07 *	0.94	0.85	-0.08 *
	36-45	0.80	0.72	-0.08 *	0.82	0.73	-0.09 *	0.96	0.86	-0.10 *
	46-60	0.84	0.73	-0.12 *	0.85	0.73	-0.12 *	0.94	0.83	-0.11
	over 60	0.92	0.85	-0.07 *	0.92	0.86	-0.06 *	0.82	0.87	0.05
White Men	18-25	0.69	0.46	-0.23 *	0.67	0.51	-0.16 *	0.71	0.69	-0.03
	26-30	0.95	0.82	-0.13	0.96	0.87	-0.08 *	0.97	0.88	-0.09
	31-35	0.94	0.86	-0.08	0.99	0.83	-0.15 *	0.99	0.75	-0.24 *
	36-45	0.98	0.83	-0.16 *	0.98	0.84	-0.14 *	0.99	0.85	-0.14 *
	46-60	0.96	0.80	-0.16 *	0.97	0.82	-0.15 *	0.98	0.86	-0.12 *
	over 60	0.93	0.93	0.00	0.94	0.87	-0.07 *	0.92	0.81	-0.11
Coloured and Indian Men	18-25	0.56	0.43	-0.14 *	0.54	0.43	-0.11 *	0.52	0.43	-0.09
	26-30	0.77	0.69	-0.09 *	0.82	0.74	-0.08 *	0.87	0.74	-0.13
	31-35	0.81	0.78	-0.03	0.84	0.82	-0.02	0.98	0.84	-0.14 *
	36-45	0.88	0.80	-0.08 *	0.90	0.82	-0.08 *	0.98	0.89	-0.08
	46-60	0.91	0.81	-0.10 *	0.91	0.82	-0.09 *	1.00	0.84	-0.16 *
	over 60	0.91	0.91	-0.01	0.91	0.90	-0.02	0.97	1.00	0.03
Black Women	18-25	0.15	0.18	0.03 *	0.14	0.20	0.06 *	0.28	0.40	0.12 *
	26-30	0.36	0.43	0.06 *	0.38	0.43	0.05 *	0.69	0.55	-0.14 *
	31-35	0.43	0.55	0.11 *	0.49	0.56	0.07 *	0.82	0.77	-0.04
	36-45	0.52	0.61	0.09 *	0.56	0.63	0.08 *	0.91	0.85	-0.06
	46-60	0.58	0.64	0.06 *	0.60	0.65	0.06 *	0.95	0.87	-0.08 *
	over 60	0.85	0.92	0.07 *	0.85	0.92	0.07 *	0.84	0.95	0.11
White Women	18-25	0.38	0.26	-0.11	0.55	0.51	-0.05	0.76	0.64	-0.12
	26-30	0.42	0.55	0.13	0.69	0.67	-0.02	0.84	0.77	-0.07
	31-35	0.50	0.59	0.09	0.68	0.66	-0.02	0.77	0.72	-0.05
	36-45	0.57	0.54	-0.02	0.64	0.66	0.02	0.76	0.79	0.03
	46-60	0.48	0.55	0.07	0.59	0.63	0.05	0.68	0.68	0.00
	over 60	0.70	0.69	-0.01	0.70	0.66	-0.05	0.71	0.59	-0.11
Coloured and Indian Women	18-25	0.40	0.37	-0.03	0.43	0.39	-0.04	0.57	0.47	-0.10
	26-30	0.50	0.54	0.05	0.57	0.61	0.04	0.88	0.81	-0.07
	31-35	0.62	0.55	-0.07	0.65	0.62	-0.03	0.88	0.89	0.01
	36-45	0.56	0.59	0.03	0.58	0.61	0.03	0.93	0.78	-0.15 *
	46-60	0.52	0.56	0.04	0.52	0.57	0.05 *	0.85	0.68	-0.17
	over 60	0.80	0.90	0.11 *	0.79	0.91	0.11 *	0.17	0.94	0.77 *

Note: Table gives income recipiency rates for cells defined by the intersection of race, sex, and age and education ranges. Columns labelled "difference" give the coefficient on a dummy for 2000 in a pooled regression of data for both years using observations only from the specified cell. The asterisk denotes statistical significance at the 5% level, and is based on the t-ratio test using standard errors clustered at the household-year level.

Appendix Table 2. Logit Coefficients

A. Logit Conditional on Endowments and Log Income

	Male		Female	
Variable	Point Estimate	Asymptotic Standard Error	Point Estimate	Asymptotic Standard Error
Black	2.754	(0.312)	5.462	(0.328)
Coloured	0.489	(0.416)	3.258	(0.425)
Indian	2.773	(0.649)	5.186	(0.749)
Age	-0.234	(0.043)	-0.389	(0.046)
Age-Squared/100	0.245	(0.051)	0.446	(0.057)
Log Income x Black	-0.232	(0.029)	-0.505	(0.032)
Log Income x Coloured	0.006	(0.040)	-0.266	(0.043)
Log Income x Indian	-0.217	(0.061)	-0.439	(0.075)
Log Income x Age	0.026	(0.005)	0.043	(0.005)
Log Income x Age-Squared/100	-0.027	(0.005)	-0.049	(0.006)
Of Pension Age	-0.274	(0.620)	-0.145	(0.620)
Education	0.065	(0.003)	0.082	(0.003)
Household Size	-0.095	(0.004)	-0.040	(0.004)
Household Head	-0.004	(0.031)	0.417	(0.025)
Log Income x Of Pension Age	0.038	(0.065)	0.012	(0.068)
Log Income	-0.781	(0.098)	-0.896	(0.105)
Constant	6.427	(0.938)	6.820	(0.969)

B. Logit Conditional on Endowments

	Male		Male	
Variable	Point Estimate	Asymptotic Standard Error	Point Estimate	Asymptotic Standard Error
Black	0.686	(0.032)	0.727	(0.035)
Coloured	0.624	(0.042)	0.605	(0.044)
Indian	0.566	(0.059)	0.747	(0.072)
Age	-0.018	(0.005)	-0.017	(0.004)
Age-Squared/100	0.022	(0.006)	0.000	(0.000)
Of Pension Age	0.038	(0.067)	-0.069	(0.052)
Education	0.022	(0.003)	0.028	(0.003)
Household Size	-0.100	(0.004)	-0.042	(0.004)
Household Head	-0.111	(0.030)	0.337	(0.024)
Constant	0.117	(0.108)	-0.118	(0.102)

Note: For each logit, the dependant variable takes a value of 1 if the year is 2000 and 0 if the year is 1995. The top panel gives the results with income included (and interacted) while the bottom panel excludes income.

Appendix Table 3. Returns to Education by Race, Age, and Gender

Population Group	Age	1995	2000	Difference	N
Black Men	18-25	0.11	0.06	0.05 *	2,855
	26-30	0.14	0.10	0.04 *	3,364
	31-35	0.13	0.12	0.01	3,569
	36-45	0.12	0.13	-0.01	6,393
	46-60	0.11	0.10	0.01	5,565
	over 60	0.05	0.04	0.01	3,764
Coloured Men	18-25	0.14	0.12	0.03	889
	26-30	0.15	0.19	-0.04 *	750
	31-35	0.17	0.19	-0.02	753
	36-45	0.16	0.17	0.00	1,361
	46-60	0.11	0.14	-0.03 *	1,268
	over 60	0.07	0.08	-0.01	663
Indian Men	18-25	0.21	0.41	-0.19	214
	26-30	0.25	0.10	0.15 *	217
	31-35	0.21	0.24	-0.03	184
	36-45	0.17	0.20	-0.03	368
	46-60	0.17	0.20	-0.04	390
	over 60	0.05	0.11	-0.06	162
White Men	18-25	0.10	0.20	-0.11 *	544
	26-30	0.12	0.16	-0.04	584
	31-35	0.18	0.20	-0.01	703
	36-45	0.18	0.19	-0.01	1,476
	46-60	0.18	0.22	-0.04	1,504
	over 60	0.17	0.18	-0.01	1,235
Black Women	18-25	0.14	0.09	0.05 *	2,466
	26-30	0.16	0.12	0.04 *	2,882
	31-35	0.16	0.13	0.03 *	3,133
	36-45	0.14	0.14	0.00	5,776
	46-60	0.10	0.11	-0.02 *	5,615
	over 60	0.02	0.02	-0.01 *	5,969
Coloured Women	18-25	0.16	0.22	-0.06 *	843
	26-30	0.18	0.21	-0.03	709
	31-35	0.18	0.24	-0.07 *	674
	36-45	0.17	0.21	-0.04 *	1,224
	46-60	0.13	0.15	-0.02	1,011
	over 60	0.01	0.04	-0.03 *	889
Indian Women	18-25	0.24	0.15	0.08	159
	26-30	0.23	0.20	0.03	127
	31-35	0.12	0.22	-0.11 *	122
	36-45	0.14	0.20	-0.07	192
	46-60	0.16	0.13	0.04	185
	over 60	-0.01	0.03	-0.04	134
White Women	18-25	0.09	0.15	-0.06	509
	26-30	0.14	0.27	-0.13 *	495
	31-35	0.16	0.13	0.02	489
	36-45	0.18	0.19	-0.01	972
	46-60	0.14	0.17	-0.03	996
	over 60	0.14	0.13	0.01	1,073

Note: Table gives coefficient on education for cells defined by the intersection of race, sex, and age. Column labelled "difference" give the difference between the 1995 and 2000 coefficients. The asterisk denotes statistical significance at the 5% level for the difference. N is the number of observations in each cell. See text for more details.

Appendix Table 4. Selection-Corrected Log Income Differences by Age, Race, and Gender

		Fraction Positive Income		1995 Incomes			2000 Incomes			Change in Incomes			1995
Group	Age	1995	2000	Lower Bound	Actual	Upper Bound	Lower Bound	Actual	Upper Bound	Lower Bound	Actual	Upper Bound	Pop. Fraction
		A	B	C	D	E	F	G	H	I	J	K	L
Black Men	18-25	0.20	0.25		9.07		8.20	8.55	8.98	-0.87	-0.52	-0.09	0.10
	26-30	0.60	0.56	9.51	9.63	9.78		9.23		-0.56	-0.41	-0.28	0.05
	31-35	0.74	0.67	9.62	9.78	9.97		9.44		-0.53	-0.34	-0.19	0.04
	36-45	0.82	0.73	9.62	9.80	10.00		9.50		-0.50	-0.29	-0.11	0.07
	46-60	0.85	0.73	9.38	9.62	9.87		9.33		-0.54	-0.29	-0.05	0.05
	over 60	0.92	0.86	8.92	9.03	9.11		8.88		-0.23	-0.15	-0.04	0.03
	Any age	0.60	0.56	9.42	9.57	9.72		9.22		-0.50	-0.35	-0.21	0.34
White Men	18-25	0.67	0.51	10.23	10.51	10.82		10.21		-0.60	-0.30	-0.02	0.01
	26-30	0.96	0.87	11.07	11.16	11.27		11.05		-0.23	-0.12	-0.02	0.01
	31-35	0.99	0.83	11.26	11.47	11.66		11.26		-0.40	-0.21	0.00	0.01
	36-45	0.98	0.84	11.34	11.54	11.74		11.37		-0.37	-0.17	0.03	0.02
	46-60	0.97	0.82	11.10	11.33	11.59		11.23		-0.36	-0.10	0.13	0.02
	over 60	0.94	0.87	10.32	10.47	10.61		10.54		-0.07	0.07	0.21	0.01
	Any age	0.92	0.79	10.92	11.14	11.39		11.02		-0.37	-0.12	0.10	0.08
Coloured	18-25	0.54	0.43	9.31	9.61	9.92		9.26		-0.66	-0.35	-0.05	0.02
and Indian	26-30	0.82	0.74	9.82	9.97	10.14		9.85		-0.29	-0.12	0.03	0.01
Men	31-35	0.84	0.82	10.07	10.12	10.18		10.18		0.01	0.06	0.11	0.01
	36-45	0.90	0.82	10.00	10.17	10.33		10.08		-0.25	-0.09	0.08	0.01
	46-60	0.91	0.82	9.75	9.96	10.13		9.82		-0.31	-0.14	0.07	0.01
	over 60	0.91	0.90	9.21	9.26	9.28		9.17		-0.12	-0.10	-0.04	0.00
	Any age	0.79	0.72	9.75	9.91	10.06		9.80		-0.27	-0.12	0.05	0.06
Black	18-25	0.14	0.20		8.85		7.73	8.24	8.79	-1.12	-0.61	-0.05	0.10
Women	26-30	0.38	0.43		9.29		8.50	8.75	9.03	-0.80	-0.54	-0.27	0.05
	31-35	0.49	0.56		9.39		8.65	8.93	9.20	-0.75	-0.46	-0.19	0.05
	36-45	0.56	0.63		9.32		8.74	9.01	9.27	-0.58	-0.31	-0.05	0.07
	46-60	0.60	0.65		9.09		8.58	8.77	8.97	-0.51	-0.32	-0.13	0.06
	over 60	0.85	0.92		8.91		8.76	8.83	8.88	-0.15	-0.08	-0.03	0.04
	Any age	0.45	0.51		9.15		8.54	8.80	9.06	-0.61	-0.35	-0.09	0.38
White	18-25	0.55	0.51	10.25	10.32	10.46		10.00		-0.46	-0.32	-0.24	0.01
Women	26-30	0.69	0.67	10.64	10.68	10.74		10.64		-0.11	-0.05	0.00	0.01
	31-35	0.68	0.66	10.59	10.64	10.69		10.68		-0.02	0.04	0.09	0.01
	36-45	0.64	0.66		10.67		10.76	10.80	10.88	0.08	0.13	0.21	0.02
	46-60	0.59	0.63		10.47		10.36	10.48	10.67	-0.11	0.01	0.19	0.02
	over 60	0.70	0.66	9.60	9.73	9.83		9.73		-0.10	0.01	0.13	0.02
	Any age	0.64	0.63	10.36	10.38	10.39		10.40		0.01	0.03	0.04	0.08
Coloured	18-25	0.43	0.39	9.40	9.54	9.72		9.00		-0.71	-0.53	-0.39	0.02
and Indian	26-30	0.57	0.61		9.70		9.37	9.52	9.71	-0.33	-0.18	0.01	0.01
Women	31-35	0.65	0.62	9.59	9.67	9.77		9.68		-0.09	0.01	0.10	0.01
	36-45	0.58	0.61		9.63		9.43	9.54	9.71	-0.20	-0.08	0.08	0.01
	46-60	0.52	0.57		9.27		9.06	9.24	9.44	-0.21	-0.03	0.17	0.01
	over 60	0.79	0.91		8.95		8.77	8.91	8.95	-0.18	-0.03	0.01	0.01
	Any age	0.56	0.58		9.48		9.23	9.32	9.45	-0.25	-0.16	-0.03	0.06

Note: Each entry in the table gives a statistic for a specific cell defined by age, sex, and race. The first two columns of the table define the cell represented in the given row. For additional details, see stub to Table 7.