A bounds analysis of school completion rates in Australia

Tue Gørgens[*] Chris Ryan[†]

April 2005

Abstract: Official estimates of school completion rates in Australia increased in the 1980s, peaked in 1992, and fell immediately thereafter before stabilizing. The official estimates were a specific focus of Australian education policy. The decline caused concern at the time. We use data from the Australian Youth Survey (AYS) to gain insight into the behavior of the official estimates. The AYS suffers from nonrepresentativeness, attrition and nonresponse which means that parameters of interest are not identified. Our bounds analysis is suggestive that school completion was overstated in the official estimates at their peak. Our analysis points to repetition as a key factor in inflating the official estimates.

Keywords: Bounds, censoring, identification, nonresponse, school completion.

JEL classification numbers: C14, C13, I21, C23, J13.

[*]Address: Research School of Social Sciences, Australian National University, Canberra ACT 0200, Australia. E-mail: Tue.Gorgens@anu.edu.au.

[†]Address: Research School of Social Sciences, Australian National University, Canberra ACT 0200, Australia. E-mail: Chris.Ryan@anu.edu.au.

1 Introduction

Nonresponse comes in a variety of forms that create problems for researchers who seek to estimate population parameters from survey data. Refusal to participate in the survey, attrition from panels and item nonresponse result in sample nonrepresentativeness and induce left- and right-censoring of longitudinal data. Individually and jointly, these forms of survey nonresponse are obstacles to population parameter estimation. As pointed out by e.g. Manski (1989, 1994), Horowitz and Manski (1998, 2000) and Zaffalon (2002), parameters of interest are usually not identified in the face of missing data.

Techniques are available that provide at least partial solutions to these problems. Nonrepresentativeness may be addressed through use of appropriate weights that match observed sample characteristics to those of the population. Values for censored data items may be imputed (see e.g. Rubin, 1987). Horowitz and Manski (1998) emphasize that such approaches often involve the imposition of untestable distributional assumptions. Since analysts often disagree about which assumptions are most appropriate, Horowitz and Manski advocate a conservative "worst case" approach. Often the data are informative about the range in which the population parameters might logically lie, and these bounds can be used as a robust means of inference and facilitate consensus among analysts: the width of the bounds reflects the information available in the sample in the absence of untestable assumptions. While this approach is intuitively appealing, the bounds are unfortunately often too wide to be informative in practice (see e.g. Little and Rubin, 1999; Raghunathan, 2000; Vazquez Alvarez, Melenberg, and van Soest, 2001).

In this paper, we provide an application where the bounds are informative. The application involves an assessment of "official" estimates of school completion rates in Australia published by the Australian Bureau of Statistics (ABS). This assessment is informed by upper bounds on progression rates between Years in upper high school estimated on data from the Australian Youth Survey (AYS). The ABS estimates were, at the time, a specific focus of Australian government education policy. In the 1980s, Australian State and Commonwealth Ministers for Education agreed to a national Year 12 completion rate target of 65% by the early 1990s. The target was exceeded within the specified time frame. After 1992, the completion rate fell by almost 6 percentage points over a number of years, a decline that caused great concern and debate among policy makers and the broader community. However, by estimating bounds on student progression from Years 10 to 11 and from 11 to 12 using AYS panel data, we find cause to doubt whether the official estimates were a good measure of school completion when they peaked in the early 1990s. We conclude that the rise and fall of the official estimates probably exaggerated changes in true school completion in Australia over the period.

The remainder of the paper is structured as follows. Section 2 describes the official school completion estimates in more detail. In section 3 we describe the AYS, and in section 4 we find that its initial samples were not representative of the population. However, we argue there that this nonrepresentativeness can be remedied with the use of appropriate weights. In section 5 we discuss remaining missing data problems in the AYS, which cannot be solved by weighting, because it is not possible to construct appropriate weights. Instead we deal with this censoring through the use of bounds as described in section 6. Section 7 presents the empirical findings. Section 8 summarizes our conclusions. The appendices contain technical details.

In Australia, levels of high school are generally referred to as "years". To help distinguish school years from calendar years, we use "Years" for the former and "years" for the latter. We use the terms "uncensored" and "observable" interchangeably.

2 The official estimates

Official statistics on school completion are published by the ABS. These figures are based on a census of school students conducted on the first Friday of August each year. (August is in the middle of the school year in Australia.) These data provide the total number of students in each Year of school in any year. The data are used by the ABS to construct a crude estimate of school completion, known as the Year 12 "apparent retention rate".

The apparent retention rate is calculated as the number of students in Year 12 in year t divided by the number of students who were in the first year of secondary school in year $t - d$, where d is the number of years of secondary school (which vary across jurisdictions). Note that this is not a measure of the proportion of individuals who commence secondary school who proceed to Year 12 in the minimum time possible, because the numerator in the ABS calculation includes individuals who were not part of the original Year cohort and some individuals in the denominator may have left Australia.[1] Since the period we analyze was characterized by population growth, the ABS figures overstate school completion.

This problem is recognized by the ABS. Each year with the publication of the estimate the ABS advises caution with its interpretation because "the method of calculation does not take into account a range of factors ... these include students repeating a year of education, migration and other net changes to the school population" (Australian Bureau of Statistics, 2002, p34).

Despite these acknowledged limitations of the measure, apparent retention rates have been used as an important performance indicator of the Australian school system. Their increase was an explicit objective of government policy, shared by Commonwealth and State governments (Australian Education Council, 1991, p8). While the last decade has seen the development of a broader range of school system performance measures in Australia, proper measurement of school completion that builds on adjusted retention rates appears to remain a priority for Australian governments (Steering Committee for the Review of Commonwealth/State Service Provision, 2003, p3.31).

To facilitate our later analysis, we separate the apparent retention rate into three multiplicative components. If N_t^g denotes the number of students doing Year g in year t,

[1] Nor is it really a measure of school "completion", but only of Year 12 participation in the middle of the school year.

then

$$\frac{N_t^{12}}{N_{t-d}^{12-d}} = \frac{N_{t-2}^{10}}{N_{t-d}^{12-d}} \times \frac{N_{t-1}^{11}}{N_{t-2}^{10}} \times \frac{N_t^{12}}{N_{t-1}^{11}}. \quad (1)$$

Unfortunately, the AYS data are not informative about the first term on the right-hand side of equation (1). Hence, our analysis focuses on the progression rates from Year 10 to 11 and from Year 11 to 12, the last two terms in the equation. Therefore, our analysis has the potential to identify whether specific problems with the measurement of progression between upper high school grades might have lead to some overestimate of the overall completion rate, but not whether that rate itself was too high.[2]

3 The Australian Youth Survey

The Australian Youth Survey (AYS) was a longitudinal survey program undertaken by the Commonwealth Department of Employment, Education and Training. The universe sampled was the Australian population, excluding those living in sparsely populated areas, aged 16–19 in September 1989 supplemented with the 16-year-old (at 1 October) population in each of the years from 1990 to 1994. No additions were made to the sample after 1994, but interviews were conducted in 1995 and 1996 for those already selected.

The six samples which together form the AYS were not drawn in the same way. Sample cohorts 1 and 2 were drawn as a nationally representative sample of persons aged 15–19 years as of 1 September 1989 who did not live in remote areas. The sample was unstratified, but geographically clustered by census collectors' districts (CDs). Those aged 16–19 were first interviewed in 1989 (cohort 1), while those aged 15 were first interviewed in 1990 (cohort 2). Each of sample cohorts 3–6 were drawn two years prior to their first AYS interview. The objective was to draw four nationally representative samples of persons aged 14 as of 1 October in each of the years 1989–1992. The sample

[2]The ABS figures show that the first factor on the right-hand side ranged from about 94% to 98% over the period we analyze. In comparison, the second and third term range between 75% to 90% (see figures 1 and 2). The latter terms were responsible for over 80% of the change in the ABS retention rate between the late 1980s and mid 1990s.

was unstratified, but clustered by schools. The selected persons were first surveyed as part of the AYS sample two years after when they were aged 16, in 1991–1994.

Table 1 shows, among other things, the size of the six sample cohorts. Because the age ranges of the populations sampled varies, the sample cohorts are not directly comparable. In order to obtain comparable groups of individuals, we work with birth cohorts instead of sample cohorts. For convenience, let *mmmyy* denote the cohort born in the 12 months following month *mmm* in year *yy*. For example, Oct69 indicates individuals born between 1 October 1969 and 30 September 1970. Table 1 shows that the October birth cohorts are roughly subsets of the sample cohorts. It is also evident that the sample cohorts include some persons not in the respective target populations. This may be due to errors in drawing the AYS sample or to miscoding of birthdays in the AYS data files.[3]

4 Representativeness of initial samples

The AYS was intended to be a nationally representative survey of the young Australians who did not live in remote areas. We need to check whether that objective was achieved and whether any departure from it can be rectified.

There are several reasons why the initial AYS samples may not be representative of the relevant populations, including failure to locate persons selected for the survey and selected persons refusing to participate. In addition, there is reason to suspect that the school based samples (cohorts 3–6) may be less representative of the population than census based sample (cohorts 1–2). The school based samples had been subject to two years of an attrition-like process in which around 300 of the original students were lost from each cohort. Of specific concern, by the time subjects from the school based samples were first interviewed in the AYS, school leavers may have been underrepresented (and

[3]We have edited the raw data (including birthdays and enrollments) for obvious errors and minor inconsistencies which could be resolved using information from elsewhere in the survey. For 111 respondents schooling information was entirely missing or inconsistent. Generally we treat both enrollment and Year as censored in these cases, but where we were able to determine the last year a respondent could possibly have been in school, we treat subsequent non-enrollment as observable. No respondents have been dropped from the sample.

more so than in the census based sample).

To assess the representativeness of the initial AYS samples, table 2 shows the marginal distribution of selected demographic variables for the AYS and for the ABS Labour Force Survey.[4] The comparison is made for October in the year where the majority of the persons in each birth cohort had their first AYS interview.

The variation due to sampling noise in the AYS is evident. In addition, it is likely that survey nonresponse contributes to the skewness. For example, females are overrepresented in the AYS for most birth cohorts, which is likely to be partly due to higher nonresponse rates for males and partly to overrepresentation of girls schools in the school based samples. The variation by state across birth cohorts is puzzling and potentially important, since school enrollment rules vary by state.

From table 2 the proportion of persons enrolled in school is too high in the AYS for most cohorts, as we anticipated. Further evidence is provided in table 3, where we present estimates of school enrollment by Year based on the ABS school census, ABS population estimates, and the ABS Labour Force Survey. These numbers underestimate the number of students doing any particular Year, since part time school students and students in home schooling are excluded.[5] Table 3 shows that the AYS has too many 16-year-olds who were doing Year 10 and too few doing Years 11 and 12. The table also confirms that school participation rates are too high in the AYS relative to the population.

These comparisons reveal some systematic differences between the AYS and the two sets of ABS data, but their magnitudes are not large and they do not provide unsurmountable difficulties for our analysis. In the first place, we will analyze conditional probabilities of moving from one Year to the next, which are likely to differ much less between the data sources than the marginal figures. Second, to take account of the effect of initial sampling bias, we construct weights which allow the AYS to match the ABS

[4] Standard errors for the AYS figures can be obtained from the ordinary "$\sqrt{pq/n}$" formula, where the sample sizes, n, are given in table 1. Note that the ABS Labour Force Survey figures are also estimates subject to random sampling errors, which are likely to be nontrivial especially for school enrollment estimates.

[5] Only "official" schools are included in the census.

Labour Force statistics on age, sex and geographical dispersion. One weight is assigned to each AYS respondent, reflecting the person's characteristics at the time of the first interview.[6] We have carried out the empirical analysis both with and without weights and found the results quite similar. This supports our expectation that estimates of conditional probabilities are much less sensitive to the nonrepresentativeness of the initial AYS samples than estimates of the marginal probabilities. For brevity we present only weighted estimates in this paper.

5 Censoring

Like most panel data sets, the AYS suffers from attrition and item nonresponse. Survey attrition means that data on outcomes of interest are not observed for some individuals. We refer to this problem as right-censoring of the individual school histories (outcome censoring in the terminology of Horowitz and Manski). Right-censoring also occurs because the survey ended in 1996. Right-censoring caused by attrition cannot be considered independent in our application, as leaving the survey is likely to be related to leaving school.

For estimating progression rates between Years of high school, the limited retrospective information in the AYS and the interaction between the AYS design and the Australian school system pose additional censoring problems. At their first interview the AYS respondents were asked if they were still in school. If the answer was affirmative, the Year was asked (and other details). If the answer was negative, then the year and month they were last in school was asked as well as the Year (and other details). No further retrospective information was collected. Respondents who were in school were asked a similar set of questions in the following interview. Respondents who had left

[6]Population growth in Australia is thus reflected in the relative size of the weighted birth cohorts in the AYS, but has no effects on the estimates otherwise. Since progression rates are conditional on previous year's school status, we do not weight by school enrollment status at the time of the first interview.

were not asked about secondary school in subsequent interviews.[7]

In some jurisdictions, many 16-year-olds were in Year 11 or 12 at the time of their first interview. Since we are interested in progression from Years 10 to 11 and Years 11 to 12 between adjacent calendar years, for some individuals in the data, these variables are censored: we do not know what Year they were in previously. We refer to this problem as left-censoring of school histories (regressor censoring in the terminology of Horowitz and Manski).

Left-censoring is closely related to Year repetition. If it was known that no-one ever repeated Year 11 then it could be inferred that someone doing Year 11 in year t must have done Year 10 in $t-1$ and left-censoring would not be a problem. Left-censoring is a problem because this person could have done either Year 10 or Year 11 in $t-1$.[8] Because progression rates for students doing a given Year vary with age (older students are generally more likely to leave school), it is not appropriate to assume that left-censoring is independent in our application.[9]

As indicated in previous paragraph, we assume that someone doing Year 11 in year t must have done either Year 10 or Year 11 in $t-1$. That is, we ignore the possibility that this person could have done Year 12 in $t-1$ and dropped back a Year, or Year 9 (or lower) and skipped a Year (or more). The amount of censoring, and left-censoring in particular, is sufficiently severe that no analysis is possible if nonobservability is left completely unrestricted. To reduce the impact of nonobservability, we make three simple and reasonable assumptions: 1) no students who left come back to secondary school in

[7]There is some ambiguity in the questionnaire for respondents who were interviewed during school breaks. We have been able to determine that respondents interviewed in December who say they were currently in school doing Year g refer to the school year that ended the previous November, whereas those interviewed in February or later refer to the current school year. Of the 99 interviews which took place in January, 46 were clearly referring to the previous year, while 51 were ambiguous and 2 were invalid. We treat the latter 53 as censored.

[8]We have investigated whether it is possible to infer the enrollment for previous years using the birthday of the person and the minimum school starting age for the jurisdiction of residence. This turned out to be very difficult, however, as a very large number of students appear to start school later than the earliest possible year.

[9]Assuming that left-censoring is independent conditional on age is not useful, because left-censoring affects the entire birth cohort: either no one in an age-Year-year combination is left-censored or everyone is.

later years; 2) no students skip a Year or drop back a Year; and 3) no students repeat more than once. These assumptions are obviously not literally true, but the number of violations in the data is extremely small.[10] These assumptions are not strong enough to allow us to infer what Year a person was doing in a particular year, if that information is not available. Rather, they allow us to narrow down the Years the person could reasonably have done.

Table 4 shows that the degree of (remaining) censoring is substantial. Censoring is particularly large for the early years, since very little retrospective information is collected in the AYS. The proportion of left-censored observations (LR+L) is much larger than right-censored (LR+R) in all years. The degree of censoring is also larger for the Year 10 to 11 progression rates than for Year 11 to 12. However, since the repeat rates for Years 10 and 11 tend to be much lower than the Year 12 repeat rate, estimates of the Year 11 to 12 progression rates may be more affected.[11]

6 Bounds

Horowitz and Manski (2000) established bounds for conditional probabilities when both the outcome and the regressor may be censored in the data, but regressor censoring does not necessarily imply outcome censoring. In general, this censoring pattern may occur for several reasons. Item nonresponse is one reason. Another is invalid, inaccurate or manipulated (e.g. "top coded") regressor data. In our application, it also occurs because certain (retrospective) data for the regressor were not collected in the survey.

Let s_t^g be the event that a person, randomly selected from the AYS universe, is in

[10] Australian students who leave high school early and later decide to complete Year 11 and/or 12 will usually do this at special institutions, not at high schools. We ignore these institutions because they are not included in the ABS school census and have not played a role in the political debate which motivates our research. In the AYS data, 3 students appear to leave and come back to secondary school (we consider this inconsistent and censor all these records), 5 students skip a grade, and 9 drop back a grade.

[11] There is no systematic data collection and no official figures published on repetition patterns in Australia. Repetition of Year 12 is more common than of other secondary Years (Nicholls, 1982), though repetition of the earliest Years of primary school is also common (Stone, 1997). Year 12 repetition is generally thought to have been higher in the the early 1990s than either before or later (Morgan, 1995).

school doing Year g in year t. The quantity of interest is the conditional probability $\mathsf{P}(s_t^g | s_{t-1}^{g-1})$; that is, the probability of progressing to the next Year in year t for those who did Year $g-1$ in year $t-1$. Let o_t^g and o_{t-1}^{g-1} denote the events that s_t^g and s_{t-1}^{g-1} are observed, respectively, in the sense that is it known whether s_t^g and s_{t-1}^{g-1} happened or not. Complementary events are denoted using an overbar.

Bounds were derived by Horowitz and Manski (2000).[12] In our notation, the lower bound is

$$L(s_t^g | s_{t-1}^{g-1}) = \frac{\mathsf{P}(s_t^g, s_{t-1}^{g-1}, o_t^g, o_{t-1}^{g-1})}{\mathsf{P}(s_{t-1}^{g-1}, o_{t-1}^{g-1}) + \mathsf{P}(o_t^g, \bar{o}_{t-1}^{g-1}) - \mathsf{P}(s_t^g, o_t^g, \bar{o}_{t-1}^{g-1}) + \mathsf{P}(\bar{o}_t^g, \bar{o}_{t-1}^{g-1})}. \tag{2}$$

The upper bound is

$$U(s_t^g | s_{t-1}^{g-1}) = 1 - \frac{\mathsf{P}(s_{t-1}^{g-1}, o_t^g, o_{t-1}^{g-1}) - \mathsf{P}(s_t^g, s_{t-1}^{g-1}, o_t^g, o_{t-1}^{g-1})}{\mathsf{P}(s_{t-1}^{g-1}, o_{t-1}^{g-1}) + \mathsf{P}(s_t^g, o_t^g, \bar{o}_{t-1}^{g-1}) + \mathsf{P}(\bar{o}_t^g, \bar{o}_{t-1}^{g-1})}. \tag{3}$$

The bounds can be estimated by replacing the probabilities in (2) and (3) by sample frequencies.

To assess the effect of sampling variation, approximate confidence regions may be placed around the estimated bounds. Several approaches have been suggested. Horowitz and Manski (1998) derived the variance of the limiting distribution of their upper and lower bounds and formed joint confidence regions using Bonferroni's method. Horowitz and Manski (2000) were also concerned with obtaining a joint confidence region for a pair of lower and upper bounds, and they suggested using the bootstrap for this purpose. Instead of a confidence region for the interval between the lower and upper bounds, Imbens and Manski (2004) proposed a confidence region which covers the true value of a parameter with a fixed (asymptotic) probability. The confidence region which is of most interest in our application is a one-sided interval which covers the true parameter with a fixed (asymptotic) probability.

[12]Theorem 1 of Horowitz and Manski (2000), with the "treatment group" consisting of the entire population.

We derive the variance of the limiting distribution of the upper bound in appendix A. A 95% one-sided confidence interval for $\mathsf{P}(s_t^g|s_{t-1}^{g-1})$ can be constructed the usual way, $\left[0, \hat{U} + 1.645\sqrt{\hat{V}_{\hat{U}}/n}\right]$, where $\hat{U}$ is an estimator of $U(s_t^g|s_{t-1}^{g-1})$ and $\hat{V}_{\hat{U}}$ an estimator of the asymptotic variance of $\hat{U}$.

7 Results

To summarize, we are interested in the time series pattern of the conditional probability of progressing from one Year to the next, $\mathsf{P}(s_t^g|s_{t-1}^{g-1})$. The official ABS estimate, which we denote by $\hat{P}_{ABS}$, is N_t^g/N_{t-1}^{g-1} from equation (1). This measure is biased upward because of Year repetition and because of population growth. We use the AYS data to calculate an upper bound, $\hat{U}$, on $\mathsf{P}(s_t^g|s_{t-1}^{g-1})$. Because school leavers are underrepresented in the AYS, this measure is also upward biased.[13] We therefore consider that our analysis is skewed towards finding that the official ABS estimates at the time did not overstate true completion. We have also estimated lower bounds, but they are not useful in this application.

The results for the progression rates from Year 10 to 11 and Year 11 to 12 are shown in figures 1 and 2, respectively, and also in table 5. There are three lines in each figure. The thick dashed line is $\hat{P}_{ABS}$ and the solid line is $\hat{U}$. The third line is $\hat{U} + 1.645\sqrt{\hat{V}_{\hat{U}}/n}$, which corresponds to the upper limit of the (pointwise) 95% one-sided confidence interval for the progression rate, based on AYS data.

Figure 1 shows that the ABS estimate is below the upper bound of the Year 10 to 11 progression rate for the entire period. However, in figure 2 the estimated ABS progression rate for Year 11 to 12 lies above the upper bound estimated from the AYS in 1990, 1991 and 1992. This means that the ABS estimates for progression from Year 11 to 12 were

[13]Since the AYS respondents report the year and the month they were last in school, it is possible to assess participation in any month of the year. To maximize comparability, the AYS estimates concern participation in August and we exclude enrollment in secondary school subjects at tertiary institutions. While participation in November-December is perhaps a better indicator of school completion, it is not known whether the respondents undertook their final examinations.

outside the range of logically possible values consistent with the AYS data. The ABS estimate of the Year 11 to 12 progression rate is inside the confidence region in most years, but in 1991 the ABS estimate is above the 95% upper confidence bound. Given the "worst case" nature of the bounds and the moderate sample sizes we take this as very strong evidence against the ABS figures.

To gain further insight into the cause of the discrepancy, we use AYS data to compute crude "benchmark" point estimates of the progression rates. The benchmarks are constructed by assuming fixed repeat probabilities to alleviate left-censoring and by assuming independent right-censoring. The assumed repetition rates are 1% per annum for Year 10, 2% for Year 11 and 5% for Year 12.[14] The construction of these estimates is described in detail in appendix B. We denote these estimates by $\hat{P}_{AYS}$.

We also compute benchmark estimates using the ABS method of calculating progression rates with the AYS data. That is, we construct estimates, $\hat{P}^{*}_{AYS}$, by taking the total number of students in Year g in year t and dividing by the total in Year $g-1$ in year $t-1$, where the totals have been adjusted for censoring. Again details are provided in appendix B. Note that students in the numerator could have been in Year $g-1$ or Year g in year $t-1$, so these are really estimates of $\mathsf{P}(s_t^g)/\mathsf{P}(s_{t-1}^{g-1})$ rather than $\mathsf{P}(s_t^g|s_{t-1}^{g-1})$; the difference being that the former is inflated by repetition (students doing Year g in $t-1$ and t). This difference means that $\hat{P}^{*}_{AYS}$ will exceed $\hat{P}_{AYS}$ and the gap between them provides a measure of how much the ABS figures may have been inflated by Year repetition, given the Year repetition patterns apparent in the AYS. Note also that $\hat{P}^{*}_{AYS}$ is not affected by population growth, in contrast to the ABS estimate $\hat{P}_{ABS}$.

Finally, we have computed naive estimates which ignore all censoring, $\hat{P}^{\#}_{AYS}$. As

[14]The figure for Year 12 is based on an Australian Government estimate from 1993 (there are no official time series estimates), while the other values are based on anecdotal observation that Year 12 repetition is much more common than repetition of other levels, supported by (censored) estimates based on the AYS data. The benchmark estimates are only moderately sensitive to the assumed repeat rates for the left-censored observations. Doubling the assumed repeat rates decreases $\hat{P}_{AYS}$ by less than 1 percentage point from 1990 onwards and by slightly more before 1990. The effect on $\hat{P}^{*}_{AYS}$ is similar from 1990. Before 1990, a doubling of the repeat rates increases $\hat{P}^{*}_{AYS}$ by 1 percentage point or less. Assuming no repeats has the opposite effects.

mentioned in section 5, progression rates are substantially age dependent, with older students less likely to stay in school. This phenomenon interacts with the age range of the AYS universe and the lack of retrospective information to bias the naive estimates downward. This is especially true for progression to Year 11. The typical AYS respondent is aged 16 at the first interview, which means that he/she is likely to be enrolled in Year 11. Since it is not known whether he/she did Year 10 or Year 11 in the previous year, the bias in the naive estimates is severe, in the order of 20–30 percentage points. For progression to Year 12, the bias is less serious and we show these estimates in figure 4 below.

Figures 3 and 4 show the ABS progression rates, the upper bounds, and the two benchmark progression rates based on AYS data. In addition, figure 4 shows the naive estimates of progression to Year 12. For progression from Year 10 to 11, we consider that the AYS estimates are most reliable in the period between 1990 and 1994. The estimates for the earlier years rely on recall and are more affected by left-censoring (as can be seen by the height of the upper bound for these years). Estimates for 1994 are slightly affected by the fact that the Oct78 birth cohort is not represented in the AYS, which means that about 7–10% of the (younger) students who did Year 10 in 1993 are missing. Estimates after 1994 are unreliable, as the average age of the AYS respondents who did Year 10 after 1993 is much older than for previous years.

Figure 3 suggests the following. First, the benchmark estimates $\hat{P}_{AYS}$ and $\hat{P}^*_{AYS}$ follow $\hat{P}_{ABS}$ closely between 1990 and 1994. Second, the differences in the magnitudes of the estimates have the expected pattern: $\hat{P}_{ABS}$ is remarkably close to $\hat{P}^*_{AYS}$, and $\hat{P}^*_{AYS}$ exceeds $\hat{P}_{AYS}$. The difference between $\hat{P}^*_{AYS}$ and $\hat{P}_{AYS}$ supports the proposition that repetition adds to estimated progression rates that do not take account of individual activity in year $t-1$.[15] It appears that the contribution of other additions to the Year cohort captured in $\hat{P}_{ABS}$, such as population growth, are of a similar magnitude to the upward bias in the benchmark estimates that results from the underrepresentation of

[15]The difference is driven principally by actual repetition patterns in the AYS, not the rates imposed on the left-censored observations described earlier, since these affect less than 20% of Year 11 students and 10% of Year 12 students (see table 4).

school leavers in the AYS.

The picture of progression from Year 11 to 12 in figure 4 is different, however. Here, we consider the data from the AYS to be most reliable for progression rates from Year 11 to 12 between 1990 and 1995. (Again the 1995 figures are slightly affected by the missing Oct78 birth cohort.) In figure 4, the estimates from the AYS do not match those of the ABS as closely as in the Year 10 to 11 case. First, the gaps between $\hat{P}_{ABS}$ and $\hat{P}_{AYS}$ are greater, especially between 1990 and 1992. Second, $\hat{P}_{ABS}$ declines relatively more after 1992 than $\hat{P}_{AYS}$. It is possible that the contribution of factors that affect $\hat{P}_{ABS}$ but not $\hat{P}_{AYS}$ changed, notably population growth and the number of older age students who undertook Year 12. Third, the gap between $\hat{P}_{AYS}$ and $\hat{P}^*_{AYS}$ is greater here than for the Year 10 to 11 case, which supports the proposition that Year 12 repetition is more prevalent than Year 11 repetition.

The naive estimate of progression from Year 11 to 12, $\hat{P}^*_{AYS}$, lies very close to, but slightly below, our benchmark Year 11 to 12 progression estimate, $\hat{P}_{ABS}$. The difference mainly reflects the assumptions of independent right censoring and fixed repeat rates used in generating the benchmark estimates. By contrast, the Year 10 to 11 naive progression estimate lies substantially below the benchmark rate (off the scale of figure 3), mainly because of the interaction between censoring and age-dependencies in the progression rate. These results suggest two things. The first is that the degree of error in adopting a naive approach to estimation in the face of missing data is very much context dependent. The second is that our benchmark estimates, $\hat{P}_{AYS}$, primarily reflects the progression apparent in the AYS data, rather than the assumptions we used its calculation.

That the AYS data matched the patterns apparent in progression from Year 10 to 11 in the ABS data but not that for progression from Year 11 to 12 in these important years is consistent with the argument that the ABS estimates were not good measures of school completion at that time. Year 12 repetition, population growth and other additions to the Year cohort from individuals outside it in year $t-1$ had a substantial impact on the ABS estimates in those years. The fact that even $\hat{P}^*_{AYS}$ is above the upper bound in much of the period further points to repetition as a key factor in inflating the ABS estimates.

While our discussion of the sources of difference between $\hat{P}_{AYS}$ and $\hat{P}^*_{AYS}$ for the Year 11 to 12 progression rates is speculative, it accords with other research that adopts the alternative approach of "correcting" the ABS estimates. Ryan and Watson (2004) adjusted the published ABS retention rates by scaling the denominator of the retention rate for population growth, limiting the age range over which the calculation was made and subtracting their own estimates of Year 12 repetition. They found that the contribution of population growth, additions of older students to the Year 12 cohort and Year 12 repetition to the estimated ABS retention rate varied over time, but peaked between 1991 and 1993 when it was some 3 to 4 percentage points higher than in either the late 1980s or the mid-1990s. Overall, repetition accounted for about half of the decline in the retention rate in the 1990s.

The ABS advises caution in interpreting its annual retention estimates and, implicitly, the underlying progression rates. The results here support the proposition that these estimates were poor measures of progression and of school completion in the early 1990s, when they peaked. These results suggest that the estimated peak in school completion was too large and the subsequent estimated decline probably too great.

8 Conclusion

In this paper, we have used the framework of Horowitz and Manski (1998, 2000) to gain insight into an official, but flawed, estimate of Australian school completion published by the ABS. We estimated upper bounds on the range of possible values progression rates from Year 10 to 11 and from Year 11 to 12 could take, consistent the characteristics of the AYS data we used. We found that the official estimates of progression from Year 11 to 12 were implausibly high at the time when the official estimates of high school completion peaked.

This application shows that a bounds analysis can be very informative. In the early

1990s in Australia, the plausible path followed by the school completion rate diverged from that of the official estimate. At the time, this estimate was a specific focus of Australian government education policy. It seems likely that this poor measure of school completion received too much emphasis as a measure of school system performance in this period.

A Variance Estimates

Use hats to denote sample analogues, and let $\hat{\Delta}$ denote the difference between the sample frequency and its expectation (for example $\hat{\Delta}(s_{t-1}^{g-1}, o_{t-1}^{g-1}) = \hat{\mathsf{P}}(s_{t-1}^{g-1}, o_{t-1}^{g-1}) - \mathsf{P}(s_{t-1}^{g-1}, o_{t-1}^{g-1})$). A first-order Taylor series expansion of $\hat{U}$ yields

$$\begin{aligned}\hat{U}(s_t^g|s_{t-1}^{g-1}) - U(s_t^g|s_{t-1}^{g-1}) = &-\frac{1}{D_{\hat{U}}}\big(\hat{\Delta}(s_{t-1}^{g-1}, o_t^g, o_{t-1}^{g-1}) - \hat{\Delta}(s_t^g, s_{t-1}^{g-1}, o_t^g, o_{t-1}^{g-1})\big)\\ &+ \frac{1 - U(s_t^g, s_{t-1}^{g-1})}{D_{\hat{U}}}\Big(\hat{\Delta}(s_{t-1}^{g-1}, o_{t-1}^{g-1})\\ &+ \hat{\Delta}(s_t^g, o_t^g, \bar{o}_{t-1}^{g-1}) + \hat{\Delta}(\bar{o}_t^g, \bar{o}_{t-1}^{g-1})\Big) + o_p(n^{-1/2}),\end{aligned} \tag{4}$$

where $D_{\hat{U}} = \mathsf{P}(s_{t-1}^{g-1}, o_{t-1}^{g-1}) + \mathsf{P}(s_t^g, o_t^g, \bar{o}_{t-1}^{g-1}) + \mathsf{P}(\bar{o}_t^g, \bar{o}_{t-1}^{g-1})$ and n denotes the sample size. For any event E, define the complementary probability $\mathsf{Q}(E) = 1 - \mathsf{P}(E)$ and use the notation $\mathsf{PQ}(E)$ as short for $\mathsf{P}(E)\mathsf{Q}(E)$. It follows that the variance of the limiting distribution as

$n \to \infty$ is

$$\begin{aligned}
V_{\hat{U}} = \frac{1}{D_{\hat{U}}^2}\Big(&\mathsf{PQ}(s_{t-1}^{g-1}, o_t^g, o_{t-1}^{g-1}) + \mathsf{PQ}(s_t^g, s_{t-1}^{g-1}, o_t^g, o_{t-1}^{g-1}) \\
&- 2\mathsf{P}(s_t^g, s_{t-1}^{g-1}, o_t^g, o_{t-1}^{g-1})\mathsf{Q}(s_{t-1}^{g-1}, o_t^g, o_{t-1}^{g-1})\Big) \\
+ &\frac{\big(1 - U(s_t^g|s_{t-1}^{g-1})\big)^2}{D_{\hat{U}}^2}\Big(\mathsf{PQ}(s_{t-1}^{g-1}, o_{t-1}^{g-1}) + \mathsf{PQ}(s_t^g, o_t^g, \bar{o}_{t-1}^{g-1}) + \mathsf{PQ}(\bar{o}_t^g, \bar{o}_{t-1}^{g-1}) \\
&- 2\mathsf{P}(s_{t-1}^{g-1}, o_{t-1}^{g-1})\mathsf{P}(s_t^g, o_t^g, \bar{o}_{t-1}^{g-1}) - 2\mathsf{P}(s_{t-1}^{g-1}, o_{t-1}^{g-1})\mathsf{P}(\bar{o}_t^g, \bar{o}_{t-1}^{g-1}) \\
&- 2\mathsf{P}(s_t^g, o_t^g, \bar{o}_{t-1}^{g-1})\mathsf{P}(\bar{o}_t^g, \bar{o}_{t-1}^{g-1})\Big) \\
- &2\frac{1 - U(s_t^g|s_{t-1}^{g-1})}{D_{\hat{U}}^2}\big(\mathsf{P}(s_{t-1}^{g-1}, o_t^g, o_{t-1}^{g-1}) - \mathsf{P}(s_t^g, s_{t-1}^{g-1}, o_t^g, o_{t-1}^{g-1})\big) \\
&\times \big(\mathsf{Q}(s_{t-1}^{g-1}, o_{t-1}^{g-1}) + \mathsf{P}(s_t^g, o_t^g, \bar{o}_{t-1}^{g-1}) + \mathsf{P}(\bar{o}_t^g, \bar{o}_{t-1}^{g-1})\big).
\end{aligned} \tag{5}$$

The derivations for the lower bound are similar. The asymptotic variances can be estimated consistently by replacing probabilities with sample frequencies.

B Benchmark Estimates

The benchmark progression estimates presented in section 7 are constructed by assuming fixed repeat probabilities for the years before the first interview for each birth cohort and by assuming right-censoring is random. We produce estimates of the number of persons doing Year g in year t and their school status in year $t+1$ by birth cohort, since these roughly correspond to the sample cohorts and the majority of the first interviews occur in the same year for each birth cohort. Let F denote the first year for which there is school information about a randomly selected person. Let F_c denote the year when (the majority of) cohort c was first interviewed.

We extrapolate the school status of each cohort backwards in time, using the following assumptions.

1. Persons who were first interviewed after F_c did not repeat any Year between F_c and the year of the first interview. That is, we extrapolate their school histories back

to year F_c assuming no repetition. (This affects only very few persons.) Formally, if s_t^g happened and $F_c < t \leq F$ then s_{t-1}^{g-1} happened.

2. Persons doing Year g in year t did either $g-1$ or g in the previous year. In other words, no-one drops back a Year and no-one skips a Year. Formally, if s_t^g happened then $\bar{s}_{t-1}^h$ happened for all $h < g-1$ and all $h > g$.

3. Year- and year-specific repeat probabilities are known.

The last assumption is obviously wrong, but critical to the extrapolation exercise. Define θ_t^g as the probability of repeating Year g in year t, $\theta_t^g = \mathsf{P}(s_t^g | s_{t-1}^g)$.

The extrapolation is based on the given repeat probabilities and simple accounting relationships. Let s_t^L denote the event that person i is not in school in year t. The structure of the AYS questionnaire implies that s_t^L is virtually always observable. Our assumptions imply that the persons doing Year g in year t must in year $t+1$ either have left school, be repeating Year g, or have progressed to Year $g+1$. Formally,

$$\begin{aligned}\mathsf{P}(s_t^g) &= \mathsf{P}(s_t^g, s_{i,t+1}^L) + \mathsf{P}(s_t^g, s_{i,t+1}^g) + \mathsf{P}(s_t^g, s_{i,t+1}^{g+1}) \\ &= \mathsf{P}(s_t^g, s_{i,t+1}^L) + \mathsf{P}(s_t^g, s_{i,t+1}^g) + \mathsf{P}(s_{i,t+1}^{g+1}) - \mathsf{P}(s_t^{g+1}, s_{i,t+1}^{g+1}).\end{aligned} \tag{6}$$

The terms in the last equation represent persons who left with g at the end of year t, repeated g in $t+1$, did $g+1$ in $t+1$, or repeated $g+1$ in $t+1$. Since $\mathsf{P}(s_t^g, s_{i,t+1}^g) = \theta_{i,t+1}^g \mathsf{P}(s_t^g)$ and $\mathsf{P}(s_t^{g+1}, s_{i,t+1}^{g+1}) = \theta_{i,t+1}^{g+1} \mathsf{P}(s_t^{g+1})$, it follows that

$$\mathsf{P}(s_t^g) = \frac{1}{1-\theta_{i,t+1}^g}\Big(\mathsf{P}(s_t^g, s_{i,t+1}^L) + \mathsf{P}(s_{i,t+1}^{g+1}) - \theta_{i,t+1}^{g+1}\mathsf{P}(s_t^{g+1})\Big). \tag{7}$$

This equation can be used to solve for $\mathsf{P}(s_t^g)$ recursively for $g = 12, 11, \ldots$ and $t = F_c - 1, F_c - 2, \ldots$. To see this, note that the AYS data identifies $\mathsf{P}(s_t^g, s_{i,t+1}^L)$ for all g and t, and $\mathsf{P}(s_{i,t+1}^{g+1})$ for $t = F_c - 1$. Furthermore, since there is no Year 13, equation (7) implies $\mathsf{P}(s_t^{12}) = \mathsf{P}(s_t^{12}, s_{i,t+1}^L)/(1-\theta_{i,t+1}^{12})$ for all t. This means that all terms on the right-hand side of (7) are identified for $g = 11$ and $t = F_c - 1$. Having identified $\mathsf{P}(s_{i,F_c-1}^{11})$, equation (7) can be used to calculate $\mathsf{P}(s^{10}_{i,F_c-1})$ and $\mathsf{P}(s^{11}_{i,F_c-2})$ and so forth.

Once the Year- and year-specific enrollments have been calculated, the progression rates follow from

$$\mathsf{P}(s^g_t|s^{g-1}_{t-1}) = \frac{\mathsf{P}(s^{g-1}_{t-1}, s^g_t)}{\mathsf{P}(s^{g-1}_{t-1})} = \frac{\mathsf{P}(s^g_t) - \theta^g_t \mathsf{P}(s^g_{t-1})}{\mathsf{P}(s^{g-1}_{t-1})} \quad \text{for } t \leq F_c. \tag{8}$$

Our benchmark estimate, $\hat{P}_{AYS}$, is constructed by replacing probabilities in (8) with sample frequencies. For $t > F_c$, to account for right-censoring $\hat{P}_{AYS}$ is the sample analogue of $\mathsf{P}(s^g_t|s^{g-1}_{t-1}, o^g_t)$.

Similarly, $\hat{P}^*_{AYS}$ is the sample analogue of $\mathsf{P}(s^g_t)/\mathsf{P}(s^{g-1}_{t-1})$. For $t \leq F_c$, $\mathsf{P}(s^g_t)$ is estimated using (7). For $t > F_c$, we estimate enrollment of each cohort under the assumptions that no-one drops back a Year or repeats more than once and that right-censoring is random. These assumptions imply

$$\mathsf{P}(s^g_t) = \mathsf{P}(s^g_t|s^{g-1}_{t-1}, o^g_t)\mathsf{P}(s^{g-1}_{t-1}) + \mathsf{P}(s^g_t|s^g_{t-1}, o^g_t)\mathsf{P}(s^g_{t-1}). \tag{9}$$

Since the data generating process identifies the enrollments for $t = F_c$ as well as the progression rates for uncensored observations, $\mathsf{P}(s^g_t)$ can be calculated for $t > F_c$ by forward recursion. Note that (9) implies that the difference between $\hat{P}^*_{AYS}$ and $\hat{P}_{AYS}$ is $\mathsf{P}(s^g_t|s^g_{t-1}, o^g_t)\mathsf{P}(s^g_{t-1})/\mathsf{P}(s^{g-1}_{t-1})$; that is, the repeat rate adjusted for the size of the Year cohorts. This adjustment factor is about 0.9 in the AYS for the relevant years.

References

Australian Bureau of Statistics (2002). *Schools Australia 2002.* Canberra: Australian Bureau of Statistics. Cat. No. 4221.0.

Australian Education Council (1991). *National Report on Schooling in Australia 1991.* Melbourne: Curriculum Corporation.

Horowitz, J. L. and C. F. Manski (1998). Censoring of outcomes and regressors due to survey nonresponse: Identification and estimation using weights and imputations. *Journal of Econometrics 84*(1), 37–58.

Horowitz, J. L. and C. F. Manski (2000). Nonparametric analysis of randomized experiments with missing covariate and outcome data. *Journal of the American Statistical Association 95*(449), 77–84.

Imbens, G. W. and C. F. Manski (2004). Confidence intervals for partially identified parameters. *Econometrica 72*(6), 1845–1857.

Little, R. J. and D. B. Rubin (1999). Comment on "Adjusting for nonignorable drop-out using semiparametric nonresponse models" by D. O. Scharfstein, A. Rotnitzky, and J. M. Robins. *Journal of the American Statistical Association 94*(448), 1130–1132.

Manski, C. F. (1989). Anatomy of the selection problem. *Journal of Human Resources 24*(3), 343–360.

Manski, C. F. (1994). The selection problem. In *Advances in Econometrics: 6th World Congress*. Cambridge, UK: Cambridge University Press.

Morgan, R. (1995). National project on repeating year 12 students. Technical report, Department of Employment, Education and Training, Canberra.

Nicholls, M. G. (1982). Short term prediction of student numbers in the Victorian secondary education system. *Australian Journal of Statistics 24*(2), 179–90.

Raghunathan, E. (2000). Comment on "Nonparametric analysis of randomized experiments with missing covariate and outcome data" by J. L. Horowitz and C. F. Manski. *Journal of the American Statistical Association 95*(449), 85–87.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons, Inc.

Ryan, C. A. and L. Watson (2004). Year 12 completion and retention in Australia in the 1990s. *Australian Journal of Labour Economics 7*(4), 481–500.

Steering Committee for the Review of Commonwealth/State Service Provision (2003). *Report on Government Services 2003*. Canberra: AusInfo.

Stone, W. (1997). Exploring links between schools and families. *Family Matters 46*, 32–33.

Vazquez Alvarez, R., B. Melenberg, and A. van Soest (2001). Nonparametric bounds in the presence of item nonresponse, unfolding brackets, and anchoring. Unpublished manuscript, Tilberg University.

Zaffalon, M. (2002). Exact credal treatment of missing data. *Journal of Statistical Planning and Inference 105*, 105–122.

Table 1: Birth Cohorts and Sample Cohorts

Birth Cohort	Sample Cohort 1	2	3	4	5	6	Full Sample	Sea†	Used Sample
Oct69	1023						1023	52	971
Oct70	1342						1342	57	1285
Oct71	1546						1546	35	1511
Oct72	1402	53	1				1456	25	1431
Oct73		1411	5				1416	14	1402
Oct74		66	1129	7			1202	1	1201
Oct75		2	11	1181	3	1	1198	0	1198
Oct76				10	1076	5	1091	0	1091
Oct77					9	1109	1118	0	1118
Other	38					1	39	1	38
Total	5351	1532	1146	1198	1088	1116	11431	185	11246

The months not part of the Oct*yy* classification are 1969m9 (38 persons) and 1979m1 (1 person). The raw data have been edited for obvious errors and minor inconsistencies. †Persons in sample cohorts 1 and 2 who arrived in Australia after the year in which they turned 14 or last went to school overseas are omitted from the analysis.

Table 2: Representativeness

Cohort	Year	Age	Male	Female	NSW	VIC	QLD	SA	WA	TAS	NT	ACT	Cap†	Bal	Enr‡
							AYS								
Oct69	1989	19	49.4	50.6	29.5	30.4	15.8	10.4	8.5	3.0	0.8	1.6	68.8	31.2	2.3
Oct70	1989	18	50.0	50.0	31.5	29.2	15.2	8.8	8.9	3.4	0.9	2.1	68.2	31.8	21.6
Oct71	1989	17	52.1	47.9	34.3	26.7	16.7	9.3	7.7	2.6	1.1	1.5	61.4	38.6	61.6
Oct72	1989	16	50.2	49.8	34.0	27.9	18.3	7.3	7.3	2.9	0.6	1.7	62.4	37.6	79.1
Oct73	1990	16	51.0	49.0	32.2	27.1	17.1	8.9	8.2	3.1	1.1	2.3	61.2	38.8	81.3
Oct74	1991	16	48.8	51.2	30.3	20.8	18.8	8.8	10.0	2.1	1.4	7.7	46.4	53.6	90.7
Oct75	1992	16	47.4	52.6	34.6	20.2	17.9	9.3	10.2	2.2	1.8	3.8	46.2	53.8	89.2
Oct76	1993	16	48.7	51.3	36.5	18.1	18.8	9.3	9.0	2.3	2.0	4.0	52.4	47.6	88.0
Oct77	1994	16	48.6	51.4	34.4	19.8	19.5	9.6	8.8	2.2	1.9	3.8	50.0	50.0	87.3
							ABS♯								
Oct69	1989	19	50.5	49.5	32.8	26.6	17.3	8.3	9.5	2.7	1.0	1.8	65.0	35.0	3.4
Oct70	1989	18	50.6	49.4	33.2	26.1	17.4	8.1	9.6	2.7	1.0	1.9	63.5	36.5	18.8
Oct71	1989	17	51.0	49.0	33.5	25.7	17.7	8.0	9.5	2.7	1.0	1.9	61.2	38.8	56.9
Oct72	1989	16	51.3	48.7	33.6	25.6	17.9	8.0	9.3	2.8	1.0	1.9	60.0	40.0	79.5
Oct73	1990	16	51.3	48.7	33.4	25.6	18.1	7.9	9.4	2.7	1.0	1.9	60.1	39.9	80.2
Oct74	1991	16	51.2	48.8	33.2	25.2	18.3	8.0	9.7	2.7	1.0	1.9	60.0	40.0	84.0
Oct75	1992	16	51.3	48.7	33.1	25.0	18.6	7.9	9.9	2.7	1.0	1.8	59.9	40.1	85.3
Oct76	1993	16	51.4	48.6	33.1	24.8	18.7	7.8	9.9	2.7	1.0	1.8	59.7	40.3	84.1
Oct77	1994	16	51.3	48.7	33.1	24.7	18.8	7.7	10.1	2.8	1.0	1.8	59.5	40.5	83.3

Legend: NSW: New South Wales; VIC: Victoria; QLD: Queensland; SA: South Australia; WA: Western Australia; TAS: Tasmania; NT: Northern Territory; ACT: Australian Capital Territory; Cap: state capital city; Bal: territory or balance of state; Enr: secondary school enrollment. †Major urban area rather than capital city for AYS sample cohorts 1–2 (Oct69–Oct73). ‡Excludes 111 cases with unknown enrollment. ♯Labour force statistics for October (ABS catalogue number 6291). Note that AYS sample sizes are given in the last column of table 1.

Table 3: July School Participation Rates

Cohort[†]	Year	Age	≤ 9	10	11	12	Enr[‡]	Left	?	June	July	N
						AYS						
Jul70	1989	18	0.0	0.7	6.5	92.8	11.5	88.3	0.2			1204
Jul71	1989	17	0.2	1.0	18.0	80.8	55.4	43.8	0.9			1456
Jul72	1989	16	0.7	15.5	66.3	17.5	76.0	22.6	1.5			1512
Jul73	1990	16	1.1	26.9	63.5	8.4	81.7	16.8	1.5			1413
Jul74	1991	16	0.4	32.1	60.6	6.9	88.9	9.7	1.3			1211
Jul75	1992	16	1.2	27.8	63.8	7.2	89.7	9.3	1.1			1221
Jul76	1993	16	0.5	30.6	62.3	6.6	89.0	9.7	1.3			1125
Jul77	1994	16	0.2	24.6	69.4	5.7	88.5	11.1	0.4			1082
						ABS[♯]						
Jul70	1989	18	0.3	1.7	13.6	84.5	9.6	90.4	4.0	8.6	10.6	
Jul71	1989	17	0.2	1.9	20.6	77.4	48.6	51.4	1.4	48.9	52.2	
Jul72	1989	16	1.0	16.9	67.2	14.9	72.9	27.1	1.2	75.0	76.8	
Jul73	1990	16	1.0	17.2	67.3	14.5	75.5	24.5	1.3	75.3	75.9	
Jul74	1991	16	1.1	17.3	67.4	14.2	79.7	20.3	1.3	83.2	82.7	
Jul75	1992	16	1.2	17.4	67.3	14.2	81.3	18.7	1.2	84.6	84.0	
Jul76	1993	16	1.0	16.8	68.5	13.6	81.3	18.7	1.1	83.3	83.2	
Jul77	1994	16	0.9	16.6	69.0	13.5	80.0	20.0	1.1	81.5	80.9	

Legend: Enr: enrolled in school; Left: persons who had left secondary school before August; ?: invalid cases (AYS) and special schools and unidentified Years (ABS); June/July: proportion of civilian population enrolled in secondary school in June/July; N: sample size. [†]The Jul74–Jul77 AYS cohorts each have three months (July–September) in one sample cohort and nine months (October–June) in another. [‡]Excludes 111 cases with unknown enrollment. [♯]Columns 4–10: School enrollments (ABS catalogue number 4221.0, various years, full-time students only) divided by population sizes (ABS catalogue number 3201.0); columns 11–12: labour force statistics (ABS catalogue number 6291.0). Note that the census of school students is taken on the first Friday of August, but tabulated by the age of the students on 1 July. Of the 11246 persons in the AYS data, 1022 persons are not in the Jul70-Jul77 cohorts.

Table 4: Left- and Right-censoring

Year	Year 10 to 11			Year 11 to 12		
	LR	L	R	LR	L	R
1987	17.5	3.7	0.0	7.9	6.1	0.0
1988	18.1	3.3	0.0	8.9	8.6	0.0
1989	7.7	12.5	0.0	1.5	16.6	0.0
1990	7.4	9.9	0.6	1.1	6.6	0.9
1991	7.7	9.3	0.7	1.5	6.5	0.8
1992	7.4	9.6	0.6	1.7	6.8	0.9
1993	6.9	9.4	0.7	1.4	6.6	1.1
1994	0.4	9.1	0.7	1.1	6.6	0.9
1995	0.5	0.0	0.5	1.1	0.0	0.6

Legend: LR: left- and right-censored; L: left-censored only; R: right-censored only. The weighted number of observations is 11208 ($11246 - 38$) in each year. In the notation of section 6, the columns LR, L and R correspond to 100 times $\mathsf{P}(\bar{o}_t^g, \bar{o}_{t-1}^{g-1})$, $\mathsf{P}(o_t^g, \bar{o}_{t-1}^{g-1})$ and $\mathsf{P}(\bar{o}_t^g, o_{t-1}^{g-1})$, respectively.

Table 5: Results

Year	$\hat{P}_{ABS}$	$\hat{U}$		$\hat{P}_{AYS}$		$\hat{P}^*_{AYS}$		$\hat{P}^{\#}_{AYS}$	
		Est	SE	Est	SE	Est	SE	Est	SE
Year 10 to 11									
1987	75.5	89.0	0.65	74.9	1.47	75.9	1.47		
1988	79.3	91.2	0.58	79.2	1.18	80.6	1.17		
1989	80.2	91.0	0.65	79.1	1.18	80.8	1.17		
1990	82.9	89.8	0.70	80.2	1.18	82.8	1.27	49.4	2.48
1991	87.6	92.7	0.59	85.6	1.02	87.8	1.15	63.1	2.46
1992	88.9	94.2	0.54	86.5	1.10	88.6	1.23	68.6	2.78
1993	88.2	93.3	0.59	85.5	1.18	87.7	1.28	65.1	3.00
1994	86.8	89.0	0.94	84.6	1.19	86.3	1.25	63.8	2.83
1995	85.9	83.6	2.06	79.3	2.86	84.6	3.33	79.0	2.87
Year 11 to 12									
1987	77.7	94.1	0.61	81.2	1.75	81.7	1.74		
1988	81.1	93.1	0.58	82.1	1.21	84.8	1.19		
1989	79.9	89.0	0.90	81.6	1.19	85.4	1.19		
1990	82.9	82.7	1.14	78.9	1.34	84.1	1.51	77.1	1.45
1991	88.6	86.6	1.01	83.1	1.18	88.0	1.46	81.6	1.30
1992	89.7	88.6	0.92	85.2	1.14	91.4	1.38	84.8	1.22
1993	87.2	89.1	0.92	86.5	1.15	92.9	1.40	85.7	1.21
1994	85.4	88.1	0.99	84.9	1.33	88.8	1.49	85.1	1.37
1995	84.6	87.9	1.00	85.8	1.32	89.9	1.52	85.8	1.32

Legend: $\hat{P}_{ABS}$: progression rates based on official ABS figures; $\hat{U}$: upper bound based on AYS data; $\hat{P}_{AYS}$: benchmark estimate based on AYS data; $\hat{P}^*_{AYS}$: ABS-type estimate based on AYS data; $\hat{P}^{\#}_{AYS}$: naive estimate based on AYS data; Est: Estimate; SE: standard error of estimate (bootstrapped for $\hat{P}_{AYS}$, $\hat{P}^*_{AYS}$ and $\hat{P}^{\#}_{AYS}$).

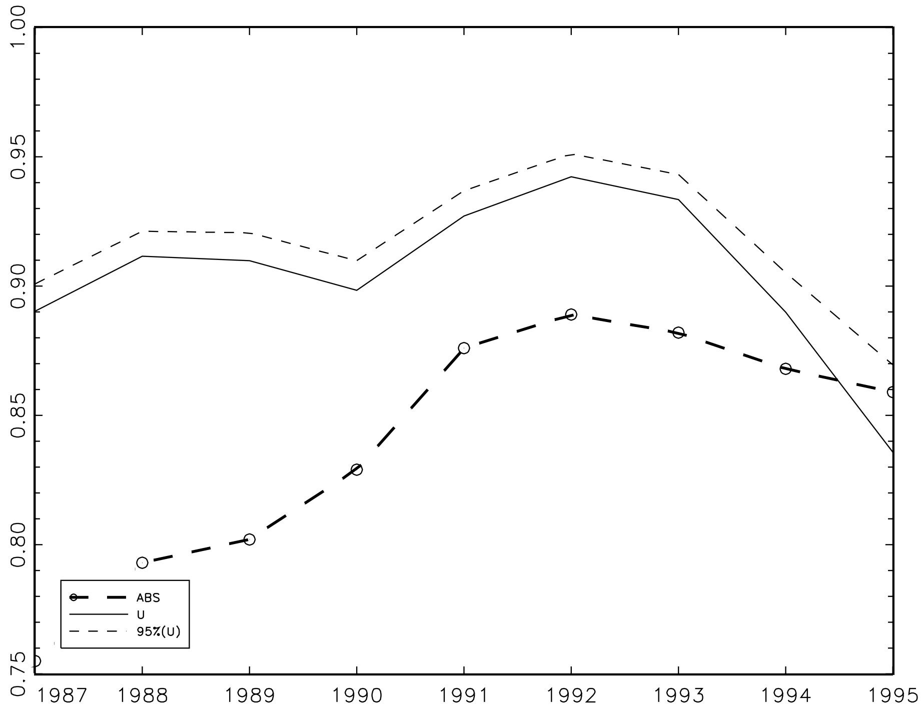

Figure 1: Year 10 to 11

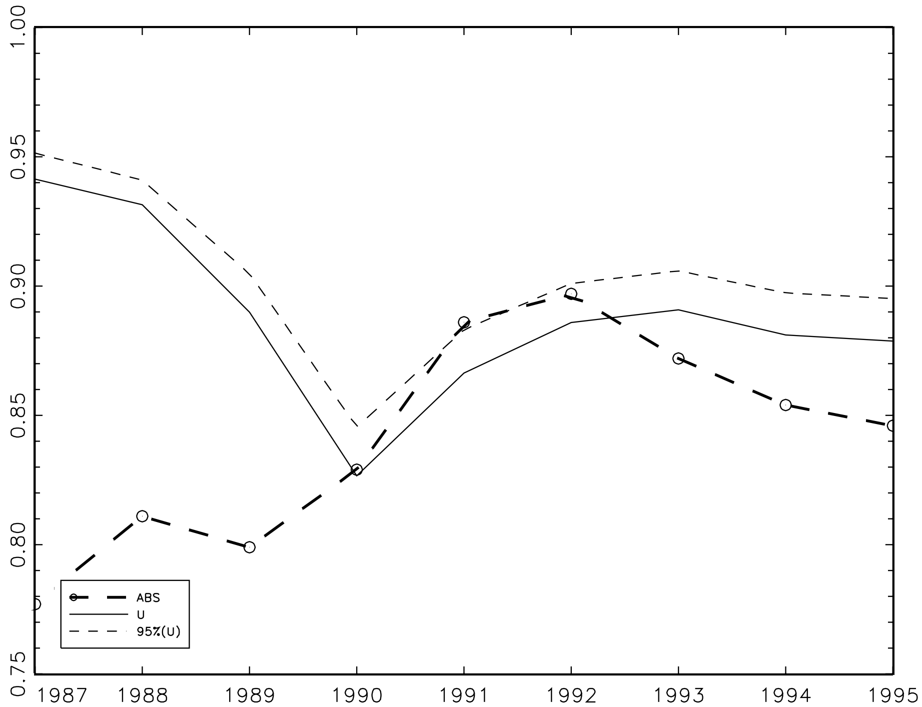

Figure 2: Year 11 to 12

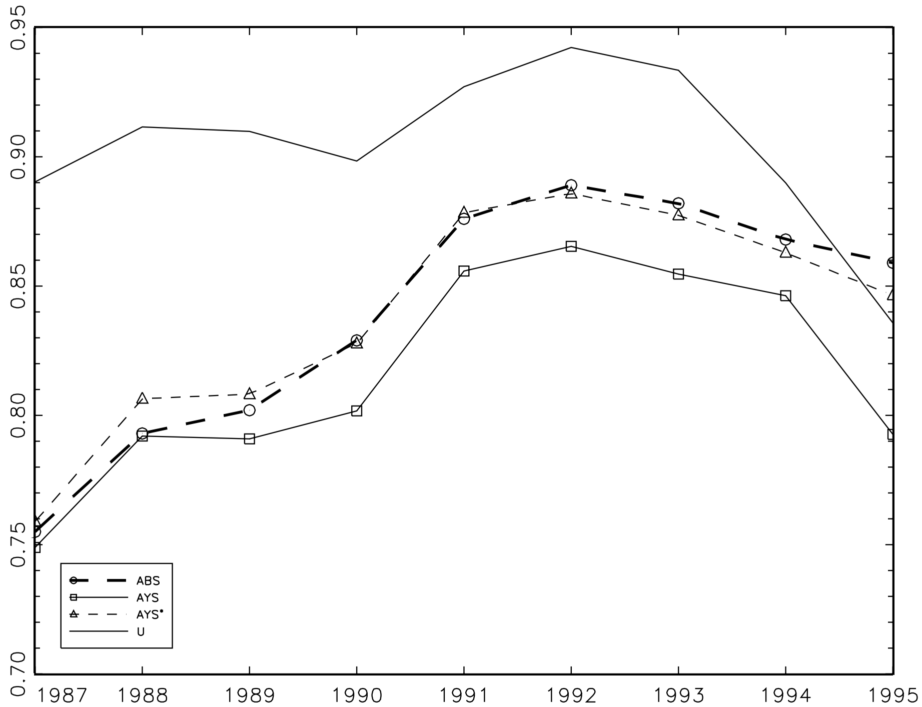

Figure 3: Year 10 to 11

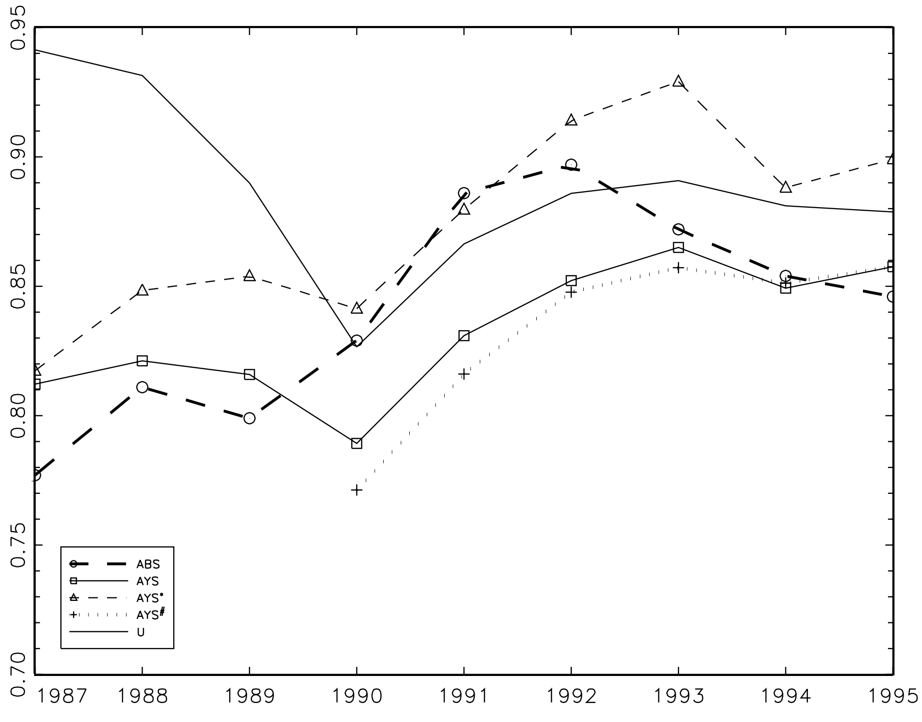

Figure 4: Year 11 to 12