Turbulence or Orderly Change?
Teacher Supply and Demand in the Age of AIDS

Luis Crouch
April 2001

An occasional paper sponsored by the Department of Education, Pretoria

This paper was commissioned by the Department of Education, Pretoria. Funding , which is gratefully acknowledged, was provided by USAID under DDSP, contract No. 674-0314-C-00-8009-00. This paper has benefited from past and ongoing work of the Department of Education and the input of many colleagues. However, the opinions and results expressed within are the sole responsibility of the author and are not to be identified with any institution with which the author works or consults. This paper was also given at the Multisite Teacher Education Research Project [MUSTER] project's Symposium on Teacher Education Policy and Practice: Insights from Research, Faculty of Education, University Pretoria, 5th – 6th April 2001.

Turbulence or Orderly Change?
Teacher Supply and Demand in the Age of AIDS

Luis Crouch[1]
April 2001
Version 1[2]

Executive Summary

This paper is constructed as a set of sections, each of which attempts to do as much exhaustive research as possible on a set of related points. However, there is little technical continuity between sections. The technical continuity, and hence the overarching logic of our argument, is provided in this executive summary only.

The author must apologise for how long it has taken to prepare this research report. Ideally, the report should have been ready in mid-2000. However, the data have proved extremely elusive; better data have been promised on an ongoing basis, and it has always seemed wise to wait for the better data. We have thus put the report on hold several times. At this point it seems wise to present the report rather than wait until the data improve even more.

Most of the analyses described in this paper are, in all likelihood, fairly boring to the reader who is not interested in methodology per se, or is not particularly interested in the specific topic of each section. This might be the majority of readers. For them, the paper itself should be treated as one long annex or footnote, with only specific sections being of particular interest; those sections and the Executive Summary are perhaps all that should be read. However, since careful methodological discussions are a good thing, the main part of the paper does focus on careful method and data analysis.

The following points summarise our arguments and provide the overall logic.

1. The most important over-arching point is that our conclusions are highly tentative, in particular in reference to teacher supply. The in-depth sociological and economic analyses of teacher identity, occupational choice, and the dynamics of the teacher labour market in South Africa, which would be needed to underpin a serious policy and planning position on these matters simply have not been done.

---

[1] Senior Economist and Director, Center for International Development, Research Triangle Institute, North Carolina, USA. Consultant to the Department of Education on finance, information, and related issues. The collaboration of numerous colleagues, particularly with the provision of data, is also gratefully acknowledged. In particular, Carol Deliwe, Rian Cilliers, Ian Bunting, and Penny Vinjevold have provided critical data that are not otherwise easily accessible. Abt Associates, in particular Saul Johnson, provided the key demographic projections data. The commissioning of the paper was by Bobby Soobrayan and Pieter Morkel, whose role in motivating and guiding the paper is acknowledged. Errors and omissions in the paper are of course attributable only to the author.

[2] Some of the data on which this study is based are shaky. In particular, the data on current enrolment in teacher training programmes in tertiary institutions are poor. As better data or formal comments come in, we will issue updated versions of this report. If no data or formal comments come in, we will not update the report. The general nature of the conclusions is unlikely to be altered, however, even with better data.

We are offering a first approximation to extremely complex issues. In our estimation, we have done at best 1/20th of the work that needs to be done before really firm conclusions about teacher identity and dynamics can be established. We challenge our colleagues and the education establishment in South Africa to undertake the necessary studies. In particular, we call for an in-depth socio-economic random sample survey of teachers and case-controls in the labour market and society at large, combined with a qualitative analysis; a study that takes the individual and collective voice of teachers seriously enough to honour it with the best research possible. We feel that the choices young people make, in terms choosing or not choosing the teaching occupation, are simply not sufficiently understood, and that unless this understanding is improved many-fold, policy and planning mistakes are very likely. Given those strong caveats we can state our conclusions.

2. South Africa's education labour market has not been as turbulent as common-sense evidence would suggest, at least in comparison with what we forecast for the future and in comparison with other countries. Entry and exit rates were actually relatively low by the late 1990s and followed relatively predictable patterns. The socioeconomic dynamics in this segment of the labour market were not markedly different from those in other sectors of the labour market. It would be hard to argue that the challenge facing the management of the system was objectively severe, yet it would seem that even then the coping capacity was in fact being tested. These assertions are demonstrated empirically in sections 2 and 5.

3. In terms of hard empirical evidence, it would be difficult to conclude that teacher compensation, relative to working hours, and compared with that of other citizens, has been unattractive in recent years—if anything, it has been ***slightly*** attractive. However, certain segments of the teacher labour market were relatively more attractive than others—in fact, some segments were actually unattractive. Being a teacher, if one is young and not well-educated oneself, is relatively attractive, but if one is well-educated and middle-aged, being a teacher is relatively unattractive. The pattern of entry and exit into teaching is shown to correspond to that relative attractiveness. Therefore, the evidence suggests that individuals were being relatively decently compensated overall, and were being rational, reacting to incentives regarding the relative unattractiveness of their segment of the labour market. In an overall sense, then, if supply is not forthcoming, it is not because individuals are acting irrationally, or because the incentives are quite poor, but because the demand is bureaucratically or budgetarily restricted, and individuals are reacting to the real probability of getting a job after exiting from teacher training. It does appear that individuals were somewhat over-reacting, in view of long-term trends. On the other hand, they were reacting fairly rationally given the limited future-oriented information at hand, and government perhaps could do more to provide more information about long-term employment prospects, based on analyses such as the present one. In some sense, the overall system is "working" as it is supposed to, but the future may require certain changes for which it may not be well prepared. These assertions are demonstrated empirically in sections 3, 4, and 5.

4. Forecasts of teacher demand and supply suggest a looming imbalance between supply and demand due in part to the AIDS epidemic, but due also in part to a) an overly hasty administrative planning process to control teacher training capacity, and b) an uncontrolled (because uninformed) and relatively short-sighted reaction on the part of young persons potentially interested in becoming teachers. However, it is only our hindsight that allows us to say this at this point. With the information that government would have had in hand in the early and middle 1990s, it is not clear it would have been logical to do anything other than what was done. And, with the information possessed by individuals in the late 1990s, their reaction in choosing not to study teaching is also understandable. Since salary incentives do not seem overwhelmingly determinant of an undersupply, and since individuals seem to be reacting rationally to incentives and information, the looming problem seems to require mostly large-scale administrative, bureaucratic, information, or planning responses, rather than wholesale reform of policy on pay levels, for example. These assertions are demonstrated empirically in section 6 in particular, but the logic is supported by sections 3 and 4 as well.

5. Our analysis suggests that addressing the AIDS-generated teacher imbalance is possible, but a real stretch. Some 30000 new teachers per year would have to be trained. Certain scenarios would make this possible. Importantly, serious forward planning of the sorts carried out in this paper would have to be engaged, and the information thus generated would have to be disseminated to secondary students who might be prospective teachers about the possibilities of finding work. Plans to insure efficient usage of capacity in teacher training programmes would have to be developed or improved. Since the cost would be enormous (some R3 billion per year), it would also be imperative to assure that such teacher training emphasises practices that make a real difference in actual learning and teaching, rather than emphasising either possible paper-chase or "feel-good" aspects which have perhaps not been uncommon in teacher training in the past. The same analysis suggests that it is unlikely to be workable for the state to attempt to pay serious and special attention to orphans through formally trained and formally paid teachers, according to the usual scales. We cannot conceive of any likely scenario that would make this feasible. The injection of some degree of informality or community-based work into the care of AIDS orphans, even at schools, is made obvious by the fact that otherwise some 50000 new formal teachers per year would be required. This seems very unlikely both because of the capital cost involved in training such teachers and the recurrent costs in paying them. The empirical documentation to back up these claims is found in section 6.

6. As noted, in spite of anecdotal evidence to the contrary, the system has not had to deal with much large-scale turbulence in recent years, and thus one could conclude that the pressure that will be put on the system is of a scale with which it is inexperienced. Thus, part of the planning response needs to be that the bureaucratic response capacity **itself** has to be improved.

7. This is all true at the macro level. But the AIDS epidemic is and will continue to be highly selective micro-regionally. Thus, part of the planning response will need to be some sort of process change to allow a better response, by the system of human resource allocation, to relative shortages at the local level. The current system is a mix of centralism (in its allocation of posts to schools and in its

determination of the processes to be followed in assigning individuals to posts), on the one hand, and individualism and "participatoriness" on the other hand in the actual assignment of individuals. This approach requires a lot of mutual checking and transacting, and it was suitable for an era of relative stability. It appears now to be on a collision course with reality, and one of the two approaches may have to be given up: either the centralism of post determination and determination of the assignment process, or the participatory and checking-intensive nature of allocation of individuals to posts. These assertions are supported empirically in section 7.

1. Background and Introduction

There seems to be little doubt in the minds of most well-informed opinion-makers that the teacher work force in South Africa has been undergoing turbulent change in the last few years. Furthermore, with the onset of the HIV/AIDS epidemic, further turbulence is predicted. Sharing this commonly held view, and wanting to put some parameters around this presumed past and future turbulence, we started out to undertake a systematic analysis of the main data sources available. We compiled the 1995, 1997, and 1999 October Household Surveys (OHS) and set up comparisons between teacher and non-teacher members of the working labour force across time. We took a cross-sectional cut of the entire PERSAL[3] database as it applied to employees of education departments in both November 1998 and November 1999 (exactly one year apart) in order to judge the dynamics of entry and exit into this database (and, hence, into the public teaching workforce) over a 1-year period. We undertook systemic demographic forward-modelling of the sector, based on the data from these sources as well as from administrative records. Finally, we analysed various other aspects of the data. We expected to be able to document great turbulence and dire trends. We expected to be able to make simple and portentous macro-level statements. What we found is worrying, but far too nuanced to drive statements that are portentous ***and*** simple.

2. Demographic Dynamics - October Household Surveys

South Africa is lucky, relative to other middle-income countries, in having a reasonably large, well-conducted, and recurrent set of household surveys that measure the socio-economic condition of the country. It may be fashionable at times, and in certain circles, to complain about StatsSA. But the reality is that the October Household Surveys , amongst other data series, are a useful source of information, sufficiently large to allow quite a lot of statistical power, and reasonably well carried out. We put together the 1995, 1997, and 1999 surveys and tried to glean what comparative data we could. There are limitations and dangers in doing this, about which more below, but we believe the risks are either worthwhile or controllable through statistical analysis—or some combination thereof. Where we are on particularly shaky ground, we warn the reader. Working with these data is not a straightforward matter of buying the CD, sticking it into one's computer, and pushing

[3] PERSAL is the central personnel database system for tracking certain characteristics of the employees of the state. It can be accessed centrally for analytical purposes, and locally for data entry.

a button. Annex C describes ***some*** of the transformations and clean-ups that were necessary, emphasising those that were needed in just one year.

An initial, tabular, foray into the data show is captured in table 1. Tabular analysis of the trends, however, is misleading, for the usual reasons plus others we will discuss below. Thus, for every single variable in the table, simple linear models were run to further assess the statistical validity of what we are about to say. These are reported in table 4. Everything below is thus stated with at least a 95% degree of confidence (though mostly with 99% confidence). In table 1, standard errors are shown, so that the reader can informally and mentally construct simple pair-wise confidence intervals. The following can be said:

- **The teaching force is approximately 20 to 25 percentage points more "feminine"** than the rest of the working labour force and is becoming more so. The entire working labour force became more "feminized" over the period 1995 to 1999. However, the teaching force did **not** become "feminized" at a faster (or slower, for that matter) rate than the rest of the working labour force. There was no statistically significant difference between the rate of feminization of the teaching force and that of the whole working labour force. There is a generally declining preference for being a teacher; if anything, women's preference for being a teacher declined faster than men's, even though the teaching force became more feminine. In other words, women were increasingly taking up opportunities in non-teaching roles faster than in teaching, at the margin.

- **Teachers say they work fewer hours per week** than the rest of the working labour force—some 17% fewer hours per week. (This does not take into account a presumably less demanding schedule over the year—we are referring only to hours per week in an actual workweek.) Furthermore, while the rest of the working labour force increased its rate of work over the period 1995 to 1997, the teaching labour force eased up its pace of work, in number of hours per week.[4]

- **Teachers report earning much higher income** than other employed persons—some 64% more. (However, they are also much more highly educated, which largely explains the higher income, as will be seen below.) Since the OHS measures income in nominal (inflation-uncorrected) terms, it is relatively meaningless to say that salaries for the overall working labour force were increasing (short of a careful analysis of price inflation relative to wage inflation).

- However, it is valid to say that **average nominal salaries for the teaching force increased faster** over the period than did those for the working labour force as a whole, than they did for the rest of the working labour force.[5] Or, that if real salaries declined for the working labour force as a whole, they declined less for teachers.

---

[4] Note that all this is by own declaration of the respondent.

[5] For the other population segments the sample size was too small to enable very firm conclusions. Also note that it is not of interest to judge whether average salaries were increasing in real terms for teachers, at least in the context of this analysis. That is why our statement is so conditional and somewhat convoluted. It is possible to judge whether the rate of change was higher for teachers than for non-teachers, but not whether it was higher or lower (for either teachers or non-teachers) than some index of price inflation.

- **Teachers are far more educated than other employed workers**—they have some 56% more years of education. Unfortunately, the operationalization of the concept "years of education" was not possible in a uniform manner from year to year in the OHS data, so time-based comparisons are not possible. We cannot tell from OHS data whether the average years of education possessed by teachers increased over the period, either in absolute terms or in comparison with other workers.

- Given their lower rate of monthly work hours, **work-hour-adjusted salaries for teachers are even higher than nominal salaries, relative to other workers**. Work-hour-adjusted salaries also grew faster for teachers than for the rest of the working labour force.

- Teachers were more unionised than the rest of the working labour force at any given point, and while the working labour force as a whole became more unionised over the period, **teachers became unionised at a faster rate than the rest of the working labour force**.

- The average working labour force age declined, but **average teachers' age increased.**

- The **most surprising factors are those relating to ethnicity or population segments**. It appears that white participation in the teaching labour force **relative** to participation in the working labour force as a whole, increased, whilst that of Africans either decreased, or stayed constant in the teaching force while it increased in the rest of the working labour force. To ensure that we are clear here, we make these statements in a rigorous, if somewhat contorted manner, since this is a surprising statement. The conditional probability of being a teacher is much higher if one is African than if one is white. But this conditional probability is decreasing for Africans and increasing for whites. Most likely this is because the rest of the formal economy is opening up at a faster rate for Africans than are opportunities in teaching for Africans, or that relative opportunities for whites have waned in the rest of the formal economy faster than in teaching, whether as a reality or as an implicit statement of preference by whites. Similarly, the conditional probability of being white, if one is a teacher, declined, whereas that of being African, if one is a teacher, increased. But the conditional probability of being white, if one is in the rest of the labour force, decreased significantly faster than did the probability of being white if one is in the teaching labour force, whereas the probability of being African, if one is a teacher, increased more slowly than did the probability of being African if one is employed in the rest of the labour force.

- These observed trends may be due to the influence of teaching posts created by School Governing Bodies (SGBs) in public schools, or even independent schools. The OHS, after all, is not restricted to teachers employed only by the public sector, but it unfortunately does not contain information to allow one to make

comparisons between publicly and privately employed teachers.[6] More on this below.

- Finally, it should be noted that this conclusion is driven by the "multivariate" analysis shown in table 4, rather than by that in table 1. The multivariate analysis can defend only the cautious statements we have made, whereas table 1 suggests less cautious statements , they are in all probability erroneous particularly when it comes to time trends, as table 4 shows. Thus, table 1 is shown only to give the reader some idea of the descriptive univariate statistics regarding these demographics. For example, the ratios pertaining to population groups in table 1 do **not** support statistically valid conclusions as to the trends we have noted.

Table 1. Characteristics of Non-Teachers and Teachers in the Working Labour Force, 1995, 1997, 1999

		Mean			Std. Error of Mean		
Year	Characteristic	Non-Teachers	Teachers	All	NonTeachers	Teachers	All
	Female	0.368	0.584	0.382	0.003	0.011	0.003
	Age	37.3	36.5	37.3	0.064	0.196	0.061
	Hours	44.2	38.8	43.9	0.070	0.211	0.068
	Union	0.333	0.554	0.349	0.003	0.011	0.003
	Toted	8.7	13.5	9.0	0.024	0.038	0.024
1995	African	0.610	0.710	0.617	0.003	0.010	0.003
	Coloured	0.122	0.075	0.119	0.002	0.006	0.002
	Indian	0.037	0.034	0.036	0.001	0.004	0.001
	White	0.231	0.181	0.228	0.002	0.009	0.002
	Salimp	2042	3403	2129	16	40	16
	Salhrsu	1971	3792	2087	18	71	18
	Female	0.379	0.608	0.391	0.003	0.013	0.003
	Age	37.4	37.0	37.4	0.070	0.239	0.067
	Hours	46.6	36.9	46.1	0.089	0.272	0.086
	Union	0.334	0.671	0.353	0.003	0.013	0.003
	Toted	8.8	13.7	9.0	0.026	0.054	0.026
1997	African	0.637	0.681	0.639	0.003	0.013	0.003
	Coloured	0.132	0.094	0.130	0.002	0.008	0.002
	Indian	0.037	0.032	0.037	0.001	0.005	0.001
	White	0.194	0.192	0.194	0.003	0.011	0.002
	Salimp	2083	3575	2171	17	65	16
	Salhrsu	1927	4084	2054	16	74	16
	Female	0.410	0.653	0.420	0.003	0.014	0.003
	Age	37.1	37.6	37.1	0.073	0.264	0.070
	Hours	46.0	37.5	45.7	0.101	0.275	0.098
	Union	0.343	0.779	0.365	0.003	0.013	0.003
	Toted	8.8	14.3	9.0	0.027	0.052	0.027
1999	African	0.652	0.681	0.653	0.003	0.014	0.003
	Coloured	0.126	0.075	0.124	0.002	0.008	0.002
	Indian	0.036	0.034	0.036	0.001	0.005	0.001
	White	0.185	0.209	0.186	0.002	0.012	0.002
	Salimp	3050	4732	3131	163	192	155
	Salhrsu	2821	5246	2939	139	208	133

Source: OHS 1995, 1997 and 1999. Author's tabulation
Note: all indicators expressed as pure ratios. Thus, 0.185, for example, is equivalent to 18.5%.

[6] This possibility was suggested by Kuben Naidoo.

As noted previously, there are problems in the sample weighting of various factors. For this reason it is less than desirable to conclude anything about the changes in a particular variable over time, without the help of other factors that could net out the effect of mere changes in weights. It is clear that if, say, the sampling weight attached to whites, or unionised persons, in the various samples was not quite right (e.g., overestimated in a particular year), then this would lead to an overestimate of the proportion of whites or unionised persons within a particular segment of the working labour force in that particular year. This could lead to the unjustifiable conclusion that (continuing the example) the proportion of whites, or of unionised persons, increased (or decreased). However, it is unlikely that overestimated weights could lead to the conclusion that, say, the proportion of whites or unionised persons in the teaching force was increasing *faster* than the proportion of the overall labour market. Whilst we may conclude erroneously—due to overestimated sampling weights—that the proportion of the teaching labour force that is white is increasing, it is very unlikely that we could erroneously conclude that the proportion of the teaching force that is white is *either* increasing faster or decreasing more slowly than the proportion that is white in the rest of the working labour force. However, it is too difficult to ascertain these phenomena based on a purely tabular analysis. Furthermore, even if these weighting problems were not present, it is simply difficult to read a table such as that presented above and draw conclusions about the simultaneous effects of time and being a teacher on the various demographic characteristics we have noted.

Reason for concern about the sampling weights is apparent from table 2. The sampling weights were normalised to average 1, separately in each year, across all employed workers. We focus on the two intersecting demographic segments at issue: whites vs. non-whites, teachers vs. non-teachers.

Table 2. Normalised sampling weights, teachers and non-teachers, whites and non-whites, 1995, 1997, 1999

	Non-teacher	Non-teacher	Non-teacher	Teacher	Teacher	Teacher	All	All	All
Year	Non-White	White	All	Non-White	White	All	Non-White	White	All
95	0.97	1.13	1.00	0.95	1.02	0.96	0.97	1.12	1.00
97	0.93	1.52	1.00	0.89	1.53	0.97	0.92	1.52	1.00
99	0.94	1.36	1.00	0.95	1.35	1.01	0.94	1.36	1.00

Source: OHS 1995, 1997, 1999; normalisation calculated by the author.

Nothing about these data would a priori necessarily indicate a problem with sampling weights, but there is something a little odd about how much the sampling weight jumps for whites. It seems a bit odd that the sampling weight for white non-teachers in 1995 is so much higher than for white teachers in the same year, and then increases to approximately the same level for both white teachers and non-teachers in 1997 and 1999. However, this may just be the way things turned out, which is fine.

The notion that the OHS data are not too biased is supported by other comparisons with the PERSAL data. Table 3 shows some key comparisons between the OHS data and the PERSAL data.

Table 3. Comparisons between OHS data and PERSAL records				
	PERSAL98	PERSAL99	OHS 97	OHS 99
Mean age	37.1	37.9	36.9	37.6
Percent female	65%	65%	61%	65%

Source: PERSAL 98 and 99, OHS 97 and 99; calculated by the author.

For this reason, and because of the simple fact that it is difficult to interpret tables for statistical significance in taking more than simple binary contrasts into account (as table 1 does), very simple models of the following types were run, for every demographic characteristic:

$$fem = a + b\,teach + c\,year + d\ \ year * teach$$

As well as

$$teach = a + b\,fem + c\,year + d\ \ year * fem$$

where *teach* and *fem* are dummy variables for being a teacher and being female, and *year* is an extremely simple time trend (0, 2, 4 for 1995, 1997, and 1999). This was run for all of the demographic characteristics listed in the table and was used to inform the bulleted points above as to the statistical significance of the various dynamics discussed. The interpretation is fairly straightforward. Taking the second equation, for example, the coefficient *b* on *fem* detects whether being female affected the probability of being a teacher in the base year, the coefficient *c* on *year* detects whether the probability of being a teacher increased over time, and the coefficient *d* on *year*fem* detects whether the passage of time increased (decreased) a female's probability of being a teacher faster (slower) than a male's. Clearly, such simple models are not meant to assess the slope or effect of the various factors but are just a quick way to test, in a "wholesale" and yet reasonably reliable manner, whether in fact there are statistically significant effects associated with some of these factors. Furthermore, these models are "multivariate" only as a technicality—which is why we use the quotation marks around the word multivariate. These models were fit using Ordinary Least Squares.[7] Note that since in many cases the dependent variable is a dummy variable, the question might arise as to the appropriateness of OLS. We tested this by running, in two cases, a logistic version of the equation above. Our finding was that it made little practical difference. The example we chose was the equation for

$$teach = a + b\,white + c\ \ year + d\,white * year$$

or, in its logistic version:

$$teach = \frac{\exp\,(a + b\ \ white + c\ \ year + d\,white * year)}{1 + \exp\,(a + b\,white + c\ \ year + d\,white * year)}$$

[7] As one would expect, the R^2s are extremely small in all these regressions. We are not seeking predictive ability, nor, much less, "causal" explanation, but rather simply to ascertain whether certain trends seem stronger than others.

A similar equation was estimated for the influence that being a union member has on the probability of being a teacher, rather than for the influence that being white has on the probability of being a teacher. Our finding was that running a simulation where the linear model was used, vs. one where the logistic model was used, led us to detect a 14.6% vs. 14.7% (*not* percentage point) increase in the estimated probability of a white person being a teacher in year 4 vs. year 0 of the 1995 to 1999 period.[8] In the case of the union dummy variable, being a union member was associated with a 131% vs. a 136% (not percentage point) increase in the probability of being a teacher in the logistic vs. linear models in 1995. These are very small differences (i.e., 14.62 - 14.57), all things considered. Given the far greater ease of interpretation of the linear model, we decided to stick with the linear versions.

Table 4 shows in summary form all of the "multivariate" relationships which we have used to double-check more rigorously, and either confirm or deny, the trends or differences implied by that table. This is done without showing the actual numbers, both to save space and because the numbers are not very important. The symbols below should be read as follows: A + implies a positive relationship, so that, for example, being a teacher implies a higher salary. A *** implies that the relationship is statistically significant at the $1/10^{th}$ of 1% level (an extremely high level of significance), ** at the 1% level, and * at the 5% level. We do not report relationships significant only at the 10% level. NS means that the relationship is "statistically not significant" (that is, more precisely, that the hypothesis that there is no relationship cannot be rejected with a reasonable level of certainty). Thus, for example, the symbols -,*** and -,*** in the columns for "Interaction of time trend and teacher dummy variable" and "Simple time trend for teachers only" and the row for "Hours worked" mean that, assuming the surveys are constructed properly, hours worked per week decreased for teachers and increased for other workers; in other words, there is only a 1 per 1000 chance that we would have measured these relationships as strongly as we did were they not true.

[8] The linear model in the case of the *white* dummy variable was:

	B	t
(Constant)	0.068	48.8
WHITE	-0.016	-5.3
YEAR	-0.005	-8.9
WH*YEAR	0.007	5.3

The logistic model was:

	B	Sig.
WHITE	-0.281	0.000
YEAR	-0.096	0.000
WH*YEAR	0.132	0.000
Constant	-2.615	0.000

Table 4. "Multivariate" analysis of simple demographic trends, using sampling weights				
Equations where the dependent variable is either a quantity (e.g., salary) or the conditional probability of being in a group				
Dependent variable	(1) Dummy variable for being a teacher	(2) Simple time trend	(3) Interaction of time trend and teacher dummy variable	(4) Simple time trend for teachers only
Salary	+,***	+,***	+,***	Irrelevant
Salary, controlling for hours worked	+,***	+,***	+,***	Irrelevant
Probability of being African	+,***	+,***	-,***	NS
Probability of being coloured	-,***	+,***	NS	NS
Probability of being Indian	NS	NS	NS	NS
Probability of being white	-,***	-,***	+,***	NS
Probability of being female	+,***	+,***	NS	+,***
Probability of being a union member	+,***	+,*	+,***	+,***
Age	-,***	-,***	+,***	+,***
Hours worked	-,***	+,***	-,***	-,***
Equations where the dependent variable is the conditional probability of being a teacher, as determined by a particular variable				
Condition or group	Dummy for the condition or group	Simple time trend	Interaction of time trend and dummy for the condition or group	
Union membership	+,***	-,***	+,***	
Female	+,***	-,***	-,*	
African	+,***	NS	-,***	
Coloured	-,***	-,***	NS	
Indian	NS	-,***	NS	
White	-,***	-,***	+,***	

Source: OHS 1995, 1997, and 1999. Author's tabulation.

Notes: 1) The fourth column in the first panel above arises out of a different equation, in each case, from that in the first three columns. In the fourth column the non-teachers have been filtered from the database and a simple time trend is fit for the relevant quantity or conditional probability. It has to be done this way because the single equation implicit in the first three columns would be over-determined if the database had been filtered for non-teachers. 2) The simple time trends for salaries for teachers only (fourth column) are irrelevant because the salary data are in nominal terms. Thus, in these two rows only the comparisons between teachers and non-teachers are relevant (third column).

To further double-check whether some of these trends are spuriously produced by weighting problems, we ran all of the above relationships without weights as well. None of the substantive conclusions changed, though some minor specifics did. (See annex A.) If anything, some of the trends look stronger if the analysis is carried out without weights. In particular, all of the time trends for population groups for teachers only (last column in the upper panel of the above table), which are NS if one uses the weights, become significant if one does not use the weights, and the

significance supports the claims we have made. For example, there appears to be a trend for whites to constitute a lower proportion of the working force altogether, but, contrary to anecdotal evidence, this trend appears weaker in the teaching force than in the rest of the employed labour force.

Above we noted that the apparent trends in the composition of the teaching force, in terms of population groups, seemed puzzling (i.e., the apparent trend for the teaching force to "Africanize" less rapidly over time than the rest of the formal labour force, as well as the difference in data, at any given point, between the OHS and PERSAL data sets). The hypothesis was presented that this may be due to the influence of independent schools and also teachers employed by SGBs in public schools (since the PERSAL data, for example, refer only to teachers employed by the public sector, whereas the OHS includes all teachers), who may be disproportionately white. It was not possible to confirm this for independent schools. However, for SGB-employed teachers, the data are very clear that SGB-appointed teachers are more frequently white than are state-paid teachers. The following discussion may throw some light on the issue.

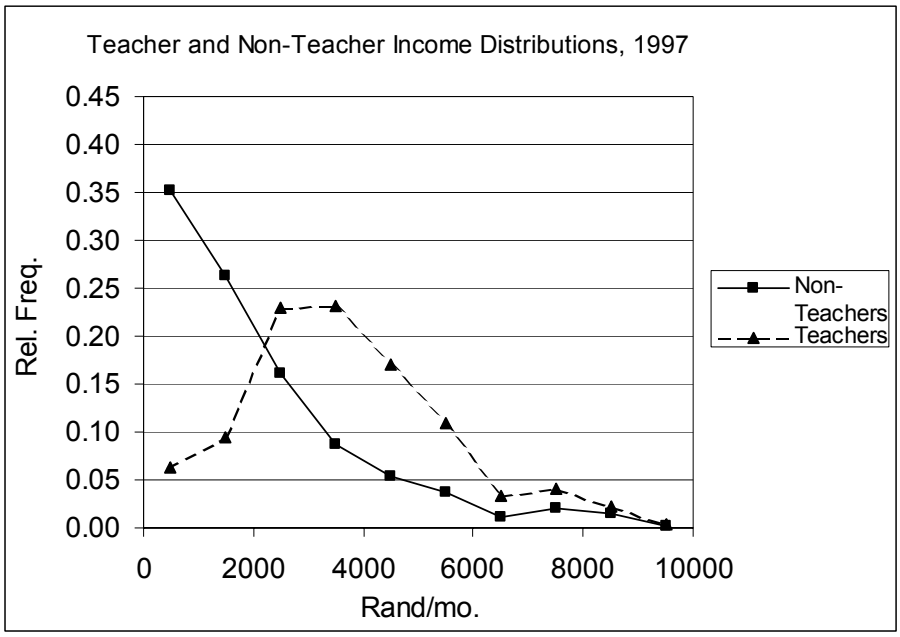
We note that according to PERSAL data, approximately 12% of teachers in the database in 1999 were white (see table 6, column 6, below). According to the OHS, however, at approximately the same time period, some 20% of teachers were white (see table 1 above). Since there were about 375000 teachers in 1998 or 1999, that 8 percentage point difference amounts to about 30000 teachers. Our best estimate, based on some sources at the Department of Education's Education Management Information System (EMIS), is that in fact some 13500 white teachers were employed in independent schools or by SGBs in public schools in 1997, and then about 25000 in 1998. This accounts for the difference between PERSAL and OHS.[9] Now, if SGB and independent school employment was growing faster than publicly paid employment in public schools during this period, as suggested by the data, then this might explain some of the dynamics noted here in this whole section. Thus, the OHS and PERSAL data are not as far from each other as one might first think. And the apparent oddity of finding a correlation between being a teacher and being white, is perhaps somewhat resolved. We would be cautious about our conclusions, however, and suggest further research on this issue.

[9] Estimated as follows using 1997 as example. According to the national Department of Education EMIS, there were about 3700 teachers employed by SGBs out of a total of some 283 000 in the relevant table. But we know there were actually about 350 000 according to EMIS itself. The reason for the difference is that the table detailing the nature of the teachers' employment was clearly not filled out by all schools. It is hard to know whether the bias would run towards schools employing a high proportion of SGB teachers (who may not wish to divulge this kind of information) or schools that employ a low proportion of SGB teachers (poor schools, which tend not to fill out the forms completely, or tend to not send their statistical returns). Without further information, we used (350/283)*3700, and applied the racial breakdown of SGB-employed teachers, to get that there must have been some 3400 white teachers employed by SGBs (70% of SGB-employed teachers were white). From EMIS data we also know there were some 14000 teachers employed by independent schools. If we apply the ratio of 70% to those, we get about 10200. We add 3400 and get about 13500, to use a round number.

3. Income Distribution of Teachers and Non-teachers

We have seen above that teachers earn higher incomes than other employed workers. We will see that much of this higher income can be explained in reference to the higher levels of education required of teachers than of other workers. In this section we want to try to develop a sense of where teachers are in the overall income distribution: are most teachers better off than, say, the top one-quarter of the non-teacher workers, i.e., the rest of the labour force? Or better off than only, say, the top one-third?

The accompanying graphic presents the basic information. The relative frequencies of earning categories are shown on the y-axis, and the earning categories are shown on the x-axis.



The income distribution for non-teachers, for example, shows that about half of non-teachers earned less than about R2000 per month (about 36% were in the R0-999 category, and about 27% were in the R1000-1999 category), whereas about half of teachers earned more than about R4000. (Note that the OHS includes independent and perhaps even relatively informal sector teachers, since it relies on the respondent's self-definition as a teacher. Thus, there are teachers earning surprisingly low salaries in the sample.)

Visually, the more the bulk of the area under the teacher's curve is displaced to the right, compared to the bulk of the area under the non-teacher's curve, the more that teachers are in the "elite" of income earners. Similarly, the smaller the y-axis value of the intersection between the two curves, the more that teachers are in an income elite relative to non-teachers (as long as the basic shape and positioning of the curves resemble those in the graphic).

Although a visual explanation may help us gain some intuition as to what we are after, it is difficult to compare many lines in one graph, particularly if they are close to one another, and yet this is what we would have to do if we wanted to compare changes over time. The following table gets at this issue by providing summary numerical measures of these factors.

Table 4. Income distribution characteristics for teachers and non-teachers

Year	Intersection of both distributions (y-axis value of the intersection)	Non-teacher percentile at the teachers' 50th
1995	0.217	80.1
1997	0.187	82.3
1999	0.172	84.3

Source: calculated by the author from OHS 1995, 1997 and 1999 data

The intersection simply shows the y-axis value at which the two curves intersect.[10] We can see that the intersection between the two curves steadily displaced itself to the right (on the x-axis) or down (on the y-axis): the bulk of teachers were shifting away from the bulk of the population, in terms of income. The other indicator shows the percentile for non-teachers at the same income level that yields the 50th percentile for teachers. Thus, it is a way of saying that, in 1995, for example, 80.1% of the employed population was worse off than the top 50% of teachers. In 1999, 84.3% of the population was worse off than the top 50% of teachers. Or, to put it another way, most (more than 50% of) teachers were better off than the top 15.7% of the working population. There was an increase of 4 percentage points in this indicator in the four years between 1995 and 1999: teachers were increasingly in the economic elite amongst the employed.[11] If we were to add the unemployed, we would see that teachers are indeed at the very top of the income distribution. Note that both

[10] This was estimated by looking at income groups on the x-axis, and the relative frequency of incomes on the y-axis. The intersection was determined by solving for both x and y in the two implicit equations for the straight line segments that intersected.

[11] There is no statistical hypothesis test that leaps to mind for ascertaining whether these trends are statistically significant. However, if we simply note that the standard error for the proportion of the population falling in, say, the 3rd income group above is about 0.006, it would appear that we can be confident at the 95% level of there being a real trend here. Since we have seen above that the average earnings of teachers did increase faster than those of non-teachers, the two results are consistent, and we can be more confident of what we are saying here. There is another problem here. Not all income data were declared as point statements of income in the OHS. Some respondents indicate simply an income category. As detailed in annexed materials, we assigned actual income to income categories, for those who did not respond with an actual income, by determining the median income for those who responded with both an income and an income category, and then assigning this within-category median to those who responded with only a category. This could create unstable borderline conditions between categories. This would be relevant if the intersection (or the non-teacher percentile corresponding to the teacher 50th percentile) falls close to one of the borders. For this reason we redid the intersection analysis above by shifting the income "bins" upward and downward by 500 Rand. The results are robust to these parametric variations, and they are available from the author.

measures drifted from each other by about 4 percentage points or 0.04 in absolute terms.[12]

In conclusion, at a macro level (though not, obviously, in the individual classroom) most teachers are better off than the parents of all but about the wealthiest 10% to 15% or so of the children they teach. It is impossible to portray teachers as belonging, as a mass, to the same socioeconomic class as the parents of the children they teach, again as a mass. And the gap seems to have increased between 1995 and 1999. Note that as countries develop, the normal trend is for teachers to start out being a relatively privileged class in society and then evolve towards being approximately in the same income/social class as most parents. South Africa's teachers are more distant from the income base of parents than is normal at her level of development. Furthermore, South Africa seems to be going slightly in opposite direction to other countries, in that as South Africa grows we are witnessing a deviation between teachers' income and that of the majority of parents. There are profound sociological dimensions to all this, and exploring them would be beyond the scope of this paper. In short, however, at this point in South Africa any identification of teachers as being in the same socioeconomic group as parents would have more to do with emotional and political issues than with economic or social ones, which is not to gainsay the importance of the former. If current trends continue, it will probably become more and more difficult to maintain whatever emotional identification parents currently have with teachers as possibly belonging to the same social groups.

4. Income Dynamics - October Household Surveys

We have noted above that teachers earn much higher incomes than other employed workers. However, they are far more educated and vary in still other ways: they are more female, more African, and more unionised than other employed workers. The question arises: do teachers have an advantage, or a disadvantage, over other employed workers in the labour force, if one considers the fact that their demographic and education profile differs from that of non-teachers? That is, does the difference in income tend to be "justified" or "explained" away when one takes into account all of the various factors simultaneously? Perhaps more importantly, if there is a "gap" in either direction between teacher and non-teacher salaries, is this gap changing, and in what ways?

This question turns out to be harder to answer than might appear at first. A logical first step is to construct a simple regression model of the determinants of income for teachers and non-teachers, using a dummy variable for teachers, assessing whether this dummy variable turns out to be an important explanatory factor, and assessing whether the size of this factor varies over time.

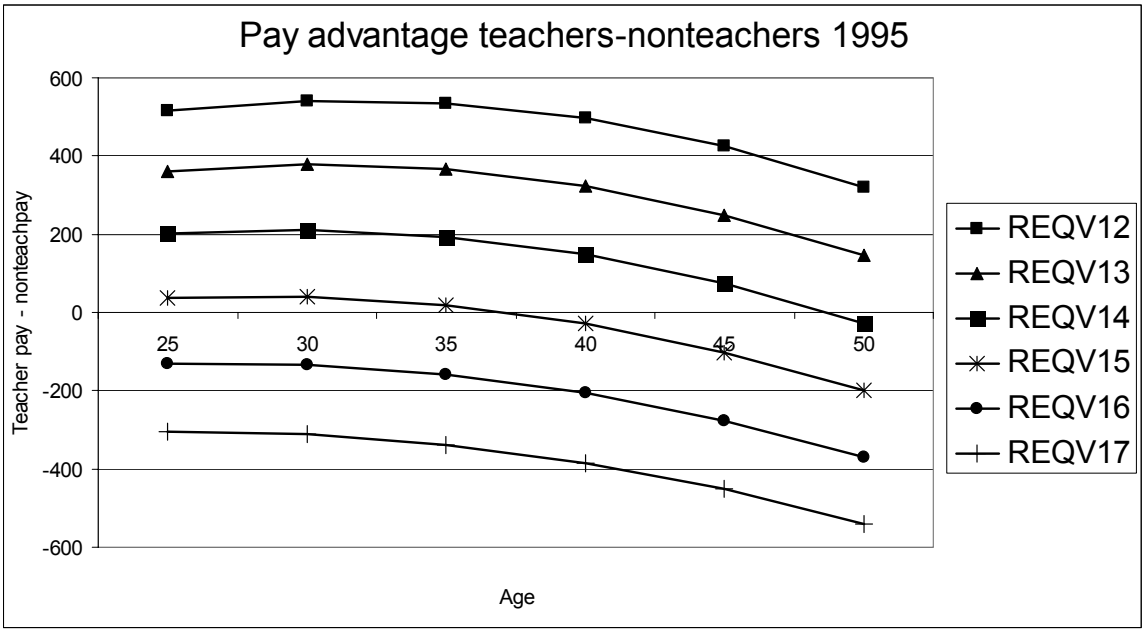
However, it turns out that much of the effect of factors such as age and education on income is nonlinear, and it is nonlinear in different ways for teachers than for non-teachers. This means that, unfortunately, one cannot simply present the regression

[12] This is logical, given that the two measures approach basically the same idea but in different ways. The first column shows the y-axis value of the intersection, whereas the second column shows the left-side area under the non-teacher's curve at the point where the left-side area under the teacher's curve covers 50% of the total area under the teacher's curve.

results and assess the significance of the teacher dummy variable. One must include many interactions of the teacher dummy variable with various linear and nonlinear versions of the education and age variables. In that case, the fact that these combinations may turn out to be important factors cannot lead to the conclusion that teachers earn more or less than is "justified" or "explainable" in reference to those factors. It may mean simply that the income of teachers and non-teachers is affected differently by those factors. To obtain the answer we seek, we must instead "simulate" the predicted income of teachers and non-teachers according to all the factors, and plot (or tabulate) these results.[13]

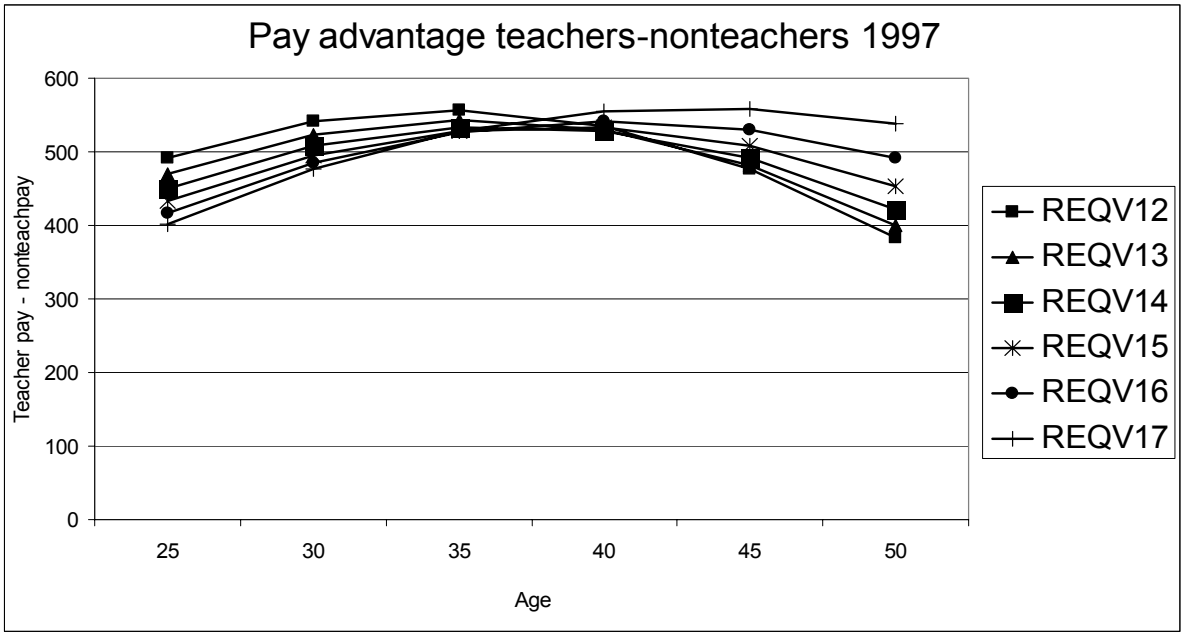
One might also wonder why one has to take a modelling approach: why not simply assess the statistical validity of the differences between teachers and non-teachers' income, using a complex tabular layout containing cells depicting age, gender, population group, and education variability. In principle this could even capture all the interaction effects. But we cannot do this because the OHS does not contain enough data in each cell to make the results interesting and meaningful. A modelling approach leverages the paucity of data in certain cells with the plenitude of data in other cells to produce results that are more easily interpretable.

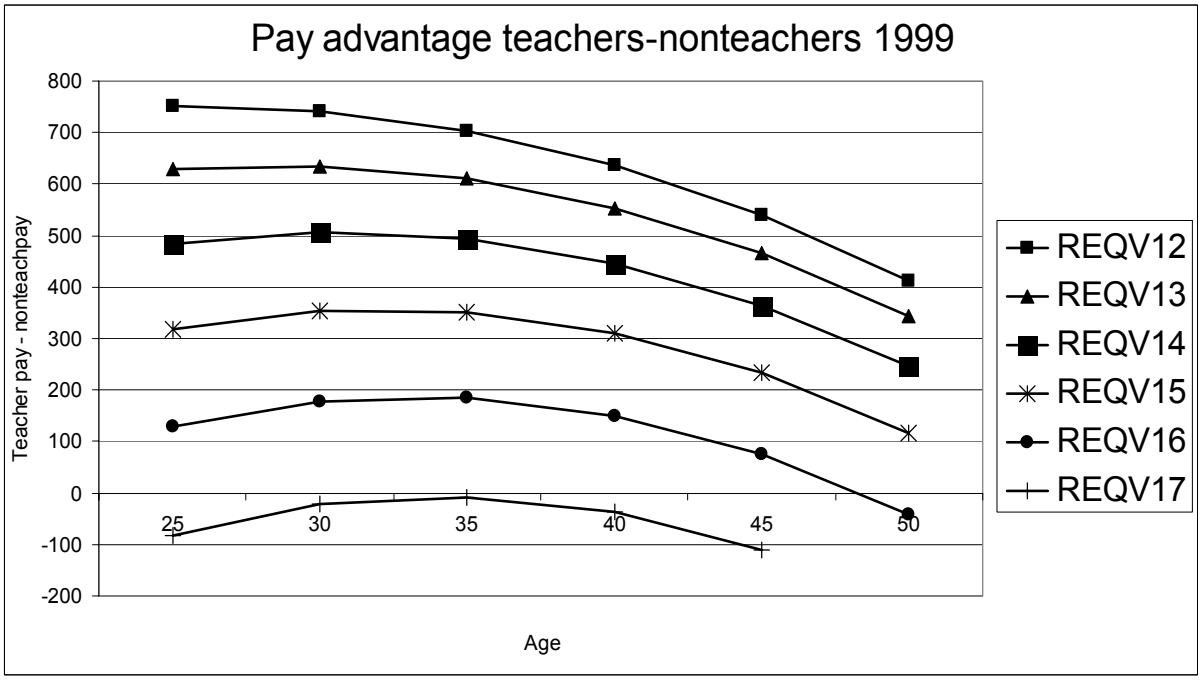
The basic regression models for the years 1995, 1997, and 1999 are presented in an annex table. Here we summarise the simulation results graphically. The simulations were calibrated to the values one would expect for an African woman belonging to a union, which is the mode demographic for teachers.[14]



[13] In fact one could take the derivative of the earnings function at various points, but this would be tedious and less informative.

[14] It would have been possible to calibrate to the mean, rather than mode, demographic for the population group, gender, union, etc., dummy variables, but not much would have been gained by doing so, at least in the version of the models we used, since there was little curvature attached to those dummy variables. The value of the intercept would have changed downward. We used the mode demographic simply to try to fix ideas on a representative individual rather than an abstract number.


Pay advantage teachers-nonteachers 1997
Teacher pay - nonteachpay
600
500
400
300
200
100
0
25
30
35
40
45
50
Age
REQV12
REQV13
REQV14
REQV15
REQV16
REQV17


Pay advantage teachers-nonteachers 1999
Teacher pay - nonteachpay
800
700
600
500
400
300
200
100
0
-100
-200
25
30
35
40
45
50
Age
REQV12
REQV13
REQV14
REQV15
REQV16
REQV17

A few points can be made on the basis of these graphics and the relevant annex table.

First, there clearly is some curvature to these relationships, and the curvature is statistically significant. The curvature with respect to age is obvious from the graphics. The curvature with respect to REQV can be noted in that the age curves for each REQV either overlap, as in 1997, or are at different widths from each other according to different REQVs, as can be seen neatly in 1999.

Second, overall, teachers in general do appear to have a pay advantage over non-teachers, even considering that they are more educated, and so forth, and ***not*** considering the fact that they work fewer hours. If we took into consideration the fact that they work fewer hours, the pay advantage would loom even larger. We evidently cannot say anything about working conditions. It is possible that working conditions for teachers are so much worse than for equally educated persons in the rest of the labour force that their pay differential is totally explainable in these terms. As a general rule, teachers seem to enjoy a relative pay advantage in the range of several hundred Rand per month that is hard to explain via reference to the various sociological and demographic factors.[15]

Third, the pay advantage in two of the studied years generally tends disproportionately to favour the young and the less educated. In 1995 fairly large groups of the older and more educated actually were compensated at rates lower than the rest of the labour force. By 1999 this had been corrected to the point where only one or two education-age groups were paid less than the equivalent segments in the rest of the labour force. The year 1997 presents perhaps the most interesting case, in that the curves actually peaked in middle age, rather than in the very young age groups, and pay for older teachers appears to have been worst for the less-educated older teachers, relative to the rest of the labour force.

Fourth, note that where older and more educated teachers do not face a pay disadvantage, as in REQV segments 14 and 15 for age groups 40 and 45 in 1999, for example, the relative pay advantage does decline with age and education—and declines at an accelerating rate. That is, the curve for REQV 15, for example, slopes downward faster between ages 40 and 45 than between ages 35 and 40. This was true in the 1995 and 1999 data, but not in the 1997 data.

---

[15] As this is being written, the press is full of discussion of teacher "poaching" by, e.g., the UK and Canada. It should be noted that the analyses of relative attractiveness and unattractiveness of the teaching labour market in South Africa, as presented in this paper, is relative to other segments of the South African labour market, not relative to teacher labour markets in wealthy countries. It should be clear to the reader that there is no way compensation policy in a country such as South Africa can be used as the main, or even simply as ***an*** important lever to retain staff. It is quite simply impossible to compete, on a purely financial basis, with the financial attractiveness of being a teacher in, say, Canada or the UK. Other factors need to be brought to bear and to be remembered, such as working conditions, the overall attractiveness of South Africa as a place of residence for skilled workers, etc. Alternatively, other "brain-drain" policies would have to be put in place to deal with such issues. (Such as ensuring that migrating individuals pay for the cost of their training if their training was publicly provided if the migration is permanent, as a sort of ex-post fee policy to pay for the cost of teacher training—obviously this would need to be researched, because if the migration is only temporary, the migration could well be a net benefit to South African education.)

Fifth, the effects of these pay issues on teachers' labour market behaviour are, at first blush, difficult to interpret. It is interesting to note, however, that the age-specific rates at which teachers abandoned teaching in the period 1998–1999 is a fairly good mirror image of the rates at which they faced a pay advantage (or disadvantage) in 1997. The curvature of the pay advantage is matched, but in mirror image, by the curvature of leaving rates. We will look into this further when we look at the rates of entry into the teaching force, below.

5. Demographic Dynamics - PERSAL

The PERSAL data allow a much more fine-grained examination of the dynamics of the teaching force. Unfortunately, unlike the OHS data, they do not allow for direct comparisons between the (public) teaching force and the rest of the (public and private) working labour force.

Perhaps the system has indeed stabilised as of 1998, and there was much more turbulence between, say, 1996 and 1998 than between 1998 and 1999, the only period we can ascertain with the PERSAL data at hand. In any case, the picture that emerges from a comparison of the numbers and characteristics of those disappearing from the database between 1998 and 1999 and those appearing in it, hardly suggests a system in turmoil, at least at the macro, national level.

Table 4 shows the leaving and joining rates, plus the sum of the two as an index of turnover (e.g., 2.6% + 0.7% = 3.3%) by province. Note that the national level is not very high, but there is enormous variance between provinces.[16]

---

[16] South Africans may find these rates high. Indeed, some early readers of this document commented on how high these rates seemed. But, as the following quotes from the USA, Canada, and New Zealand show, South Africa's rates of turnover, at least during this period, are low, or at worst on par with even the developed countries:

> 1. "A recent report by the Professional Practices Commission to the State Board of Education showed that almost 20 percent of all first-year teachers leave the teaching profession. Two primary reasons for the high attrition rate are a lack of compensation and inadequate training, according to the report." (*The Chronicle Online*, chronicle.duke.edu/chronicle/95/11/01/01StateReport.html). The statment refers to North Carolina only. Similar quotes can be produced in quick order for most American states.
>
> 2. "New teachers, especially those tossed into "sink or swim" situations, often do not survive in "real world" classrooms. Nationally, it is estimated that 30 percent of new teachers leave during their first two years, and more than 40 percent depart during their first four years. Studies also show that teachers who leave the profession reported a lower mean income than those who stayed, challenging the belief that teachers quit to earn more money in other careers." (Mentoring and Leadership Resource Network, mentors.net/LibraryFiles/OutaHere.html). Note—see below—similarity of these rates for young teachers between these U.S.-wide numbers and those of South Africa.
>
> 3. "On a Canada-wide basis, 21.7% of teaching staff had left their jobs in the previous 12 months. Of those who left, 38.1% quit voluntarily, 13.3% were fired for poor performance, 11.5% were laid off for reasons such as decreased enrollment or their time-limited contract period ended, and 11.0% took a leave of absence. A leave of absence was most frequently associated with maternity or parental leave. The remaining 26.1% of staff who left did so for a variety of un-stated reasons." (Child and Family Canada Website, http://www.cfc-efc.ca/docs/00001054.htm). Note to South Africa: 13.3% fired for non-performance.

Table 5. Turnover characteristics of various provinces and national level between 1998 and 1999

	Leaving	Joining	Turnover ratio	Net change
EC	2.6%	0.7%	3.3%	-1.9%
FS	6.5%	1.4%	7.9%	-5.0%
GT	8.6%	3.9%	12.5%	-4.6%
KN	6.3%	0.8%	7.1%	-5.6%
MP	3.5%	3.3%	6.7%	-0.2%
NC	6.8%	3.2%	10.0%	-3.6%
NP	6.0%	1.4%	7.4%	-4.6%
NW	2.8%	0.8%	3.6%	-2.0%
WC	5.8%	5.5%	11.3%	-0.3%
National	5.3%	1.9%	7.3%	-3.4%

Source: calculated by the author from PERSAL database.
Notes. 1. Numbers do not always add up perfectly due to rounding error.
2. The turnover ratio used above is not the standard human resources turnover ratio. We use the sum of the "leaving" ratio and the "joining" ratios, whereas the standard measure is the "leaving" ratio. Given the low correlation between the "leaving" and "joining" ratios in this case, we felt that the sum is more indicative of total movement.

The wealthiest and traditionally best-endowed provinces (Western Cape, Northern Cape, and Gauteng) underwent considerable turnover, whereas in the poorest provinces the turnover was lower. Naturally, there is a net loss from the sector (e.g., 5.3% - 1.9% = 3.4%). The numbers for net loss follow no clearly discernible patterns.

Table 5 shows the basic demographic characteristics of those who were in the database in 1998, those who apparently left it, and those who joined it.

---

4. "Attention also needs to be given, however, to teacher turnover and mobility at the micro level and to the relative capacity of schools in particular settings to recruit and retain good quality teaching staff. Generally tightness or shortfalls in total supply are most sharply felt in schools which have been difficult to staff on a historic basis… It is more likely that the effect of such disincentives contributed to the markedly higher turnover rate of staff in the 79 rural primary schools surveyed by the Office. In these schools the number of vacancies in the preceding calendar year calculated as a percentage of the total teacher workforce in those schools was in excess of 32 percent. That labour turnover rate was 12 percent higher than the average of primary schools located in cities and almost six percent higher than schools located in provincial towns." (Education Review Office, New Zealand, www.ero.govt.nz/Publications/eers1996/96no8hl.htm).

Table 6. Basic demographic characteristics of leavers and joiners from the teaching force, as compared to the entire teaching force 1998 to 1999						
Characteristic	Leavers as a % of those with same characteristic in 1998 database	Joiners in 1999 as a % of those with same characteristic in 1998 database	Net leavers as a % of those in the 1998 database	Leavers with the given characteristic as a % of all leavers	Joiners with the given characteristic as a % of all joiners	% of those with the given characteristic in the 1998 database
Gender						
Male	5.37%	1.79%	3.58%	34.8%	32.4%	34.6%
Female	5.33%	1.97%	3.36%	65.2%	67.6%	65.4%
Population group						
African	4.2%	1.6%	2.6%	60.5%	64.4%	76.2%
Coloured	4.8%	2.9%	1.9%	7.5%	12.6%	8.3%
Indian	7.5%	1.4%	6.1%	4.5%	2.3%	3.2%
White	12.0%	3.2%	8.7%	27.5%	20.7%	12.3%
REQV						
10	9.1%	1.6%	7.5%	6.5%	4.2%	3.6%
11	5.8%	0.3%	5.5%	4.6%	0.7%	4.0%
12	3.8%	0.2%	3.6%	11.8%	2.5%	15.5%
13	4.6%	1.7%	2.9%	34.8%	46.1%	38.6%
14	6.2%	1.9%	4.2%	31.3%	36.1%	25.8%
15	4.5%	0.8%	3.7%	8.1%	5.3%	9.1%
16	4.3%	0.5%	3.9%	2.5%	1.0%	3.0%
17	6.5%	0.8%	5.7%	0.4%	0.2%	0.3%

Source: calculated by the author from 1998 and 1999 PERSAL database.
Note: we have not presented standard errors in order to minimise overload on the table. However, in general, all of the implicit differences above are statistically significant. For example, the differences between leaving and joining rates for males and females are statistically significant, as is the difference between joining rates at REQV 13 and 14.

The PERSAL data confirm the OHS results that the teaching force is not just female-dominated, but increasingly so.[17] More men than women left, and more women than men joined. And, the proportion of women joining is larger than the proportion of women in the base.

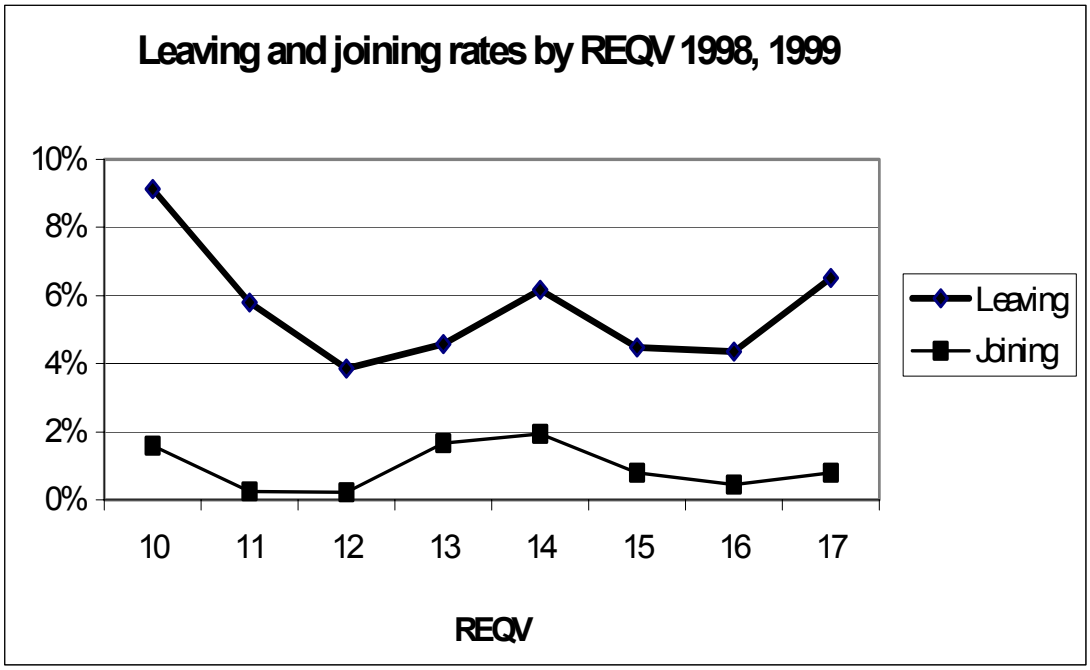
The data by population group are interesting but also puzzling. First, as already noted, these data contradict the OHS data. We saw that the OHS suggests that about 20% of all teachers are white, as recently as 1999. However, the PERSAL database suggests that only some 12% of teachers are white. It is quite possible that this is largely due to teachers being employed by school governing bodies (SGBs) and by independent schools. It would require that there be some 30000 white teachers employed by independent schools, SGBs, or in informal situations. We discussed this phenomenon above.

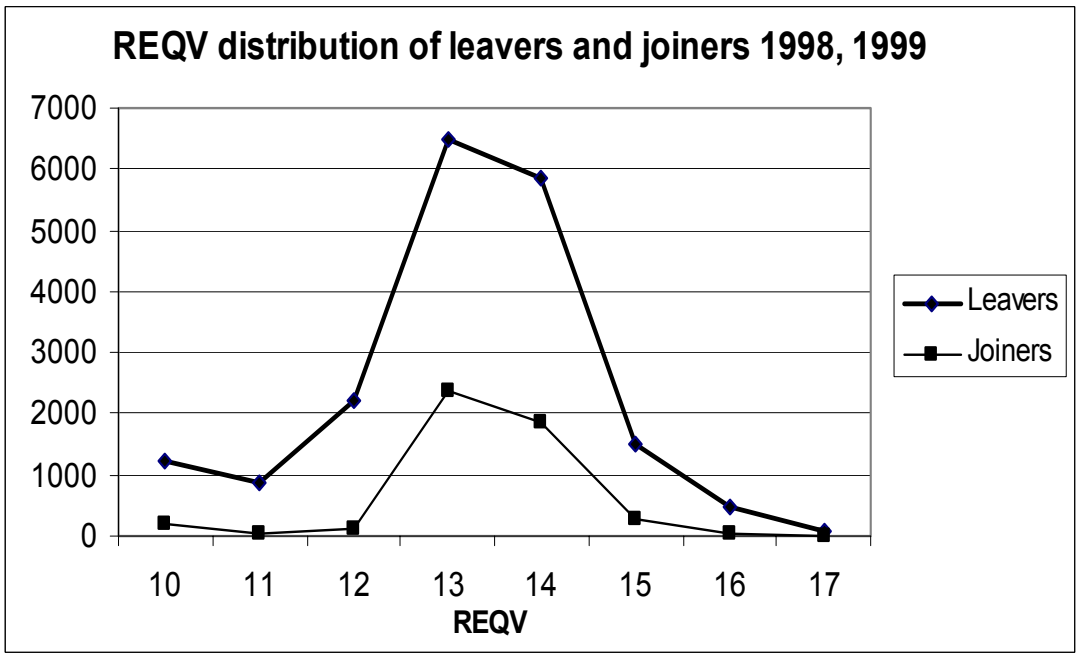
In any case, aside from this issue, the data are still of interest. It is obvious that Africans, for example, are leaving ***and*** joining the teaching force in proportions smaller than they are in it. While some 76% of teachers are African, 61% of those

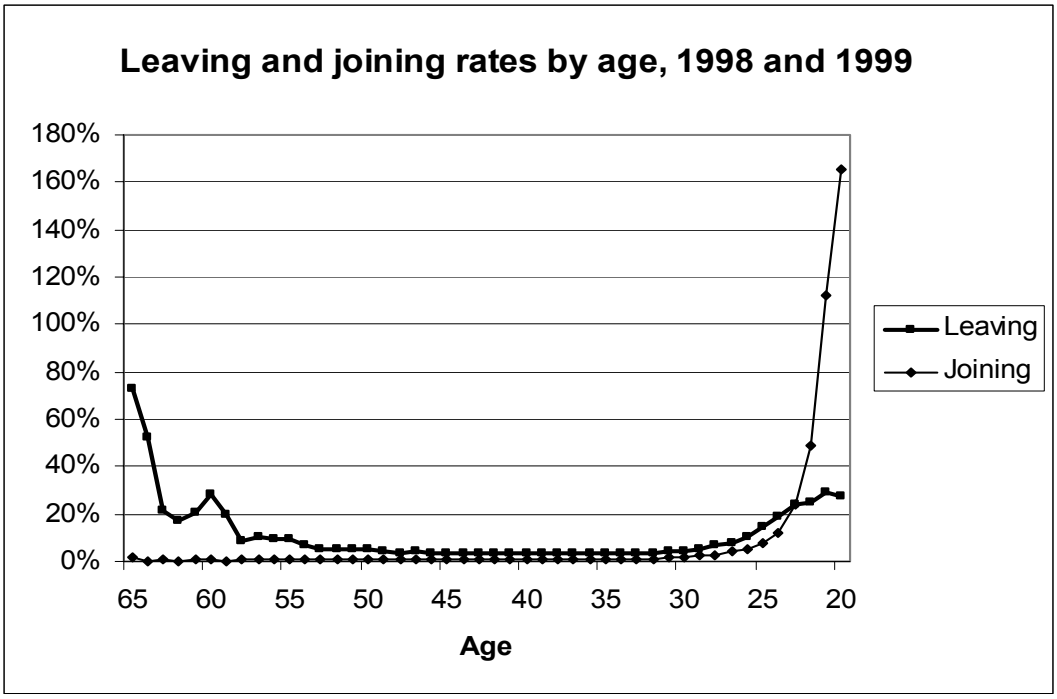
[17] Even though with PERSAL we are dealing with administrative records where the sample therefore equals the universe, all of our statements have been assessed for statistical significance, and we discuss only results that are valid with at least a 95% confidence interval.

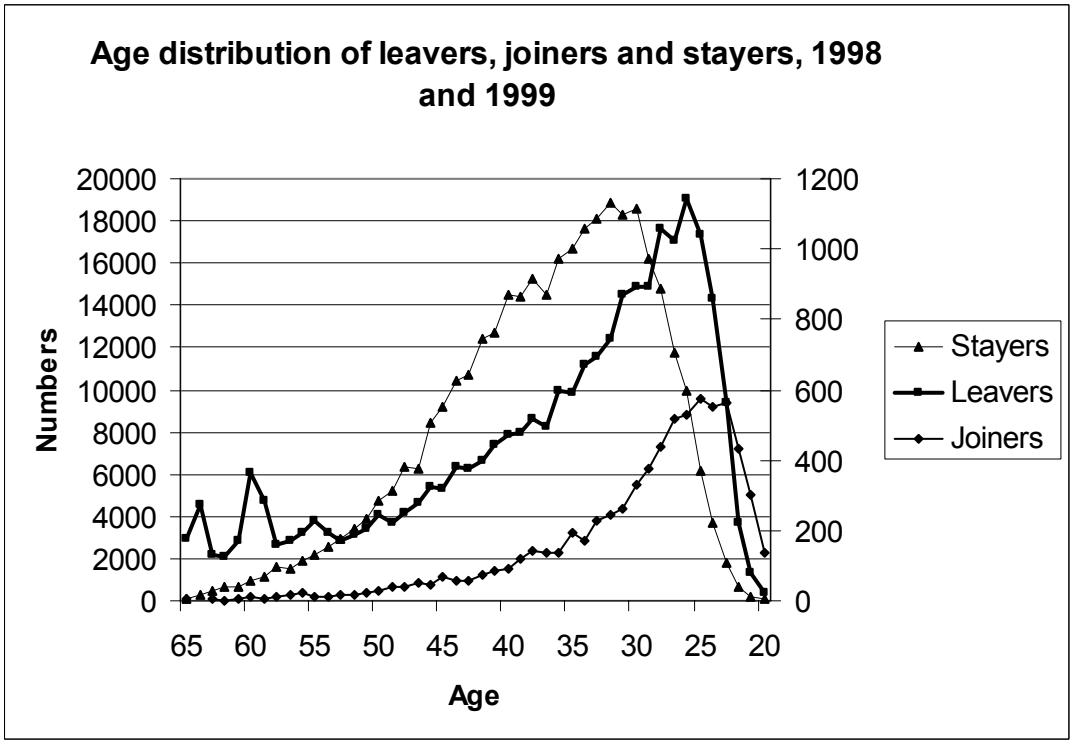
leaving are African, and 64% of those joining are African. On the opposite side, while some 12% of teachers are white, about 28% of those leaving are white, but, interestingly, some 20% of those joining are also white. This evidently deserves some further examination. It is possible that there is a sort of "churning" of white teachers out of and back into the teaching force. The dynamics for coloured teachers were the most extreme, in that they are joining at rates much greater than their proportion of the teaching force. These factors suggest if not "turbulence," at least interesting dynamism.

On the other hand, an examination of the PERSAL data shows remarkable predictability in the age distribution of both those joining the database and those leaving it. The following four graphics illustrate the issues quite aptly. Note that for each concept (age and REQV) there are two graphs: the leaving and joining rates, and the distribution of leavers and joiners. The former gives one a sense of how the "average" teacher of a given age or REQV is reacting and behaving, whereas the latter gives a sense of how that behaviour translates into numbers leaving and joining. For example, we see below that teachers with REQV 14 were leaving at a faster rate than were teachers with REQV 13, as a proportion of the total number of REQV 14 teachers. But, because there are more teachers with REQV 13, the total number of teachers with REQV 13 who were leaving was higher than that of teachers with REQV 14 who were leaving.


Leaving and joining rates by REQV 1998, 1999
10%
8%
6%
4%
2%
0%
10
11
12
13
14
15
16
17
REQV
Leaving
Joining


REQV distribution of leavers and joiners 1998, 1999
7000
6000
5000
4000
3000
2000
1000
0
10
11
12
13
14
15
16
17
REQV
Leavers
Joiners


Leaving and joining rates by age, 1998 and 1999
180%
160%
140%
120%
100%
80%
60%
40%
20%
0%
65
60
55
50
45
40
35
30
25
20
Age
Leaving
Joining


Age distribution of leavers, joiners and stayers, 1998 and 1999
Numbers
20000
18000
16000
14000
12000
10000
8000
6000
4000
2000
0
1200
1000
800
600
400
200
0
65
60
55
50
45
40
35
30
25
20
Age
Stayers
Leavers
Joiners

The rate of leaving (those leaving divided by those in the database) by REQV is what one would expect and corresponds to policy: those with the least education are leaving at the fastest rate. There are two other peaks: at REQV 14 and REQV 17. Interestingly, the pattern for the joining rate reflects the pattern for the leaving rate but naturally at a lower level, given that there was net outflow from the database.

If we look at the REQV distribution of the numbers (not rates) of leavers and joiners, we note a much clearer peak at REQV 13. Again, however, the REQV distribution of those leaving and those joining is extremely similar, suggesting stability in the system. In terms of the tabular analysis, the REQV data are also of some interest. The composition of those joining is much more tightly concentrated around REQVs 13 and 14 than the composition of the base (i.e., 46.1% and 36.1% of the joiners are at REQV 13 and 14, as opposed to 38.6% and 25.8% for the base). This suggests that the teaching force is becoming less diverse in terms of training.

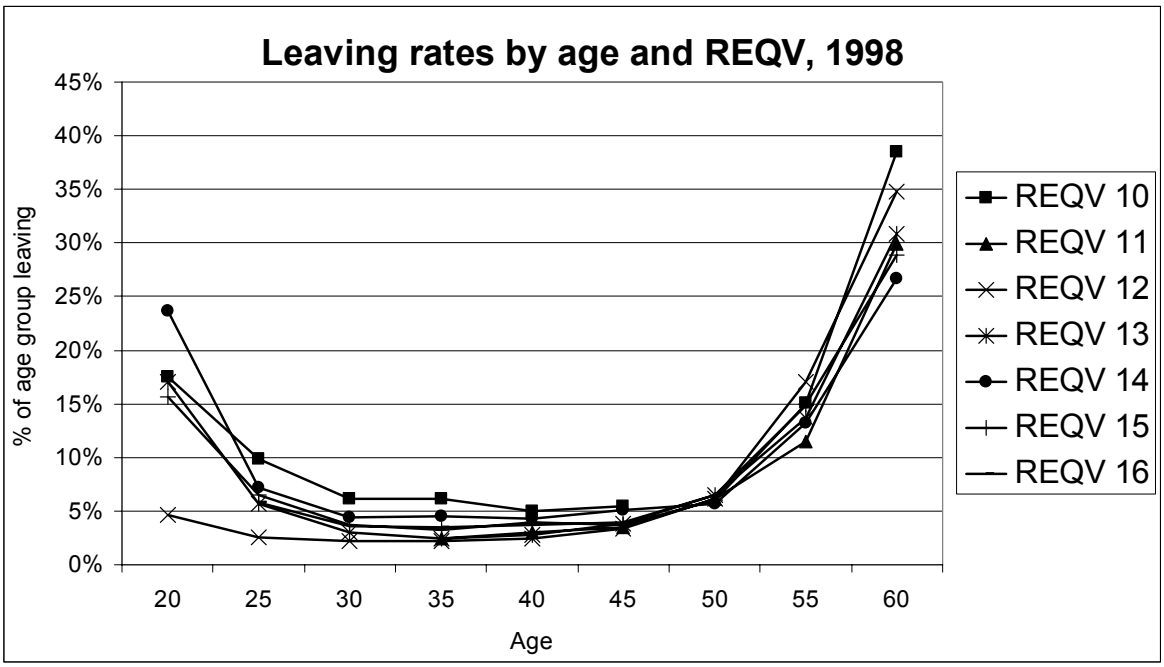
The age distribution of leavers and joiners also suggests stability and order, though with some surprises. Above all, these figures suggest that the notion of teachers leaving the profession during their (presumably most productive) middle-level years is either mythical or is no longer the case. The pattern of leaving suggests that those who leave are mostly either ready for retirement (note the peaks at exactly 65 and 60) or are very young and therefore simply giving the profession a try or joining while awaiting better prospects, and leaving quickly upon finding them. Note that in terms of the age distribution, the peaks for leaving are 65 and 21; for joining, around 20; and for those who stay, 32. Note how closely the distributions for joining and for leaving resemble each other, except for the old-age peak in the latter. Note that the average age of teachers is about 37.

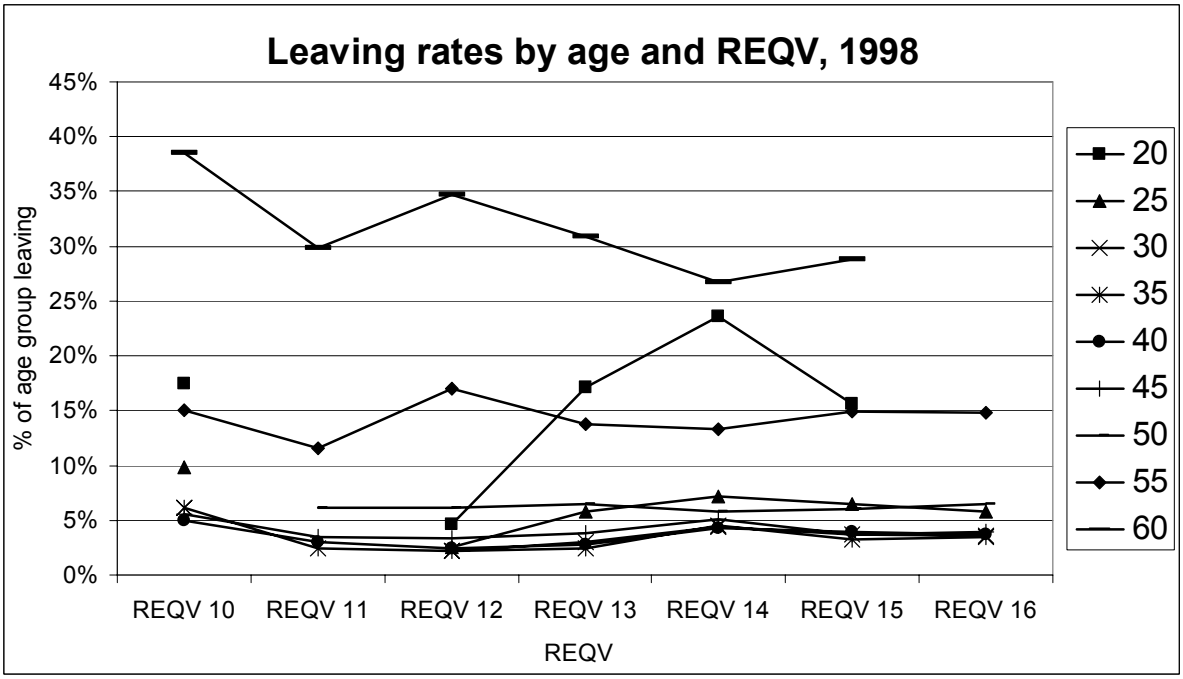
What is very interesting about both the age and REQV data is that the leavers and joiners are more like each other than they are like those who stay or who were in the database to begin with. For example, those who leave ***and*** those who join tend to be more concentrated around REQVs 13 and 14 than those who were in the teaching force at the start. Similarly, those who join and those who leave tend to be very young. That joiners would be young is expected, that the young would leave at such fast rates, and that so many of the leavers would be young, is somewhat surprising, however.

The following two graphics combine the REQV and age data. Because there are insufficient data points in each combination of REQV and single-year age groups, we utilised 5-year age groups. This is why the data appear smoother, from the x-axis point of view, in the following graphic than in those preceding. Points containing less than 25 leavers have been excluded from the graphic, as a simple statistical-validity precaution. That is why there is no curve for REQV17, and it is also why some points are missing from some of the other curves, though this is not easily visible to the naked eye. Note that the two graphics present exactly the same information, but from two different perspectives.

It is clear that teachers with different REQVS respond in extremely similar ways to the pressures (or opportunities) for leaving the profession, with the possible exception of the low leaving rate for REQV12 at the younger end of the age spectrum.

It is also clear that the bulge in those who leave at REQV 14 is concentrated amongst young teachers (the 20 and 25 age groups). Otherwise, in other age groups, leavers are concentrated amongst the less-educated teachers. This is not surprising, since it is the younger and least educated teachers who seem to have enjoyed the best pay advantage (or least disadvantage) in the mid 1990s.

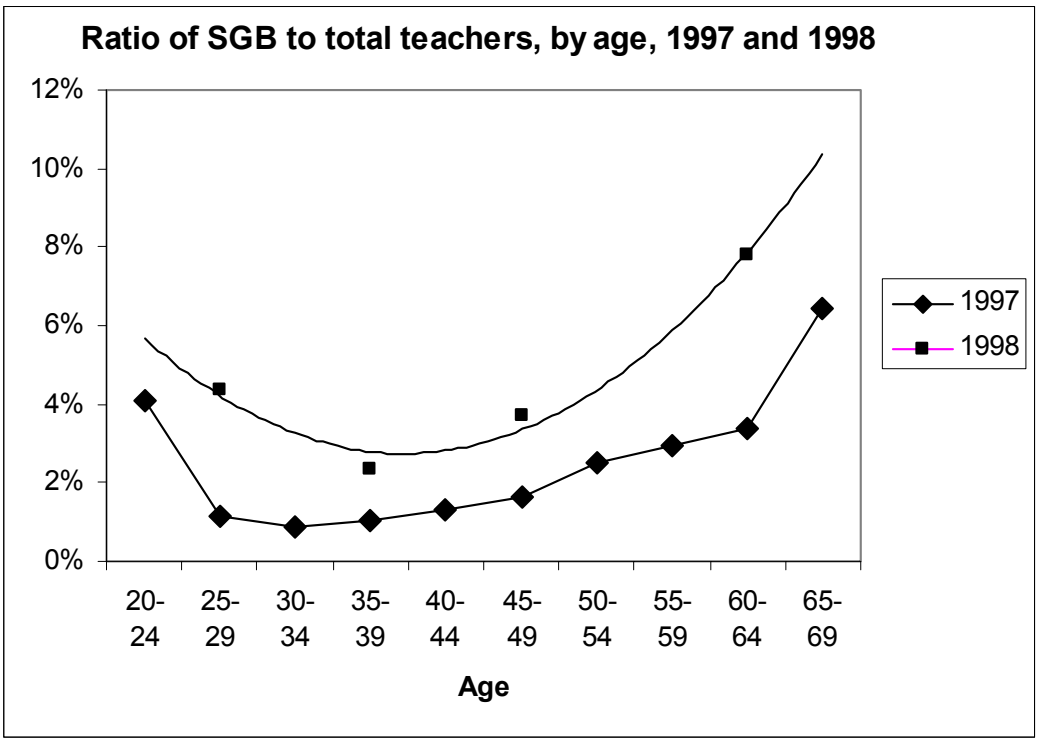




We noted above (section 2) that the relationship between the curvature in pay advantage and the rates of leaving and joining is a bit puzzling. On the whole, we noted that young and less-educated teachers have the greatest relative pay advantage over non-teachers in 1995 and 1999, whereas in 1997 the pay advantage is more to the middle-aged.

Having a pay advantage would explain the high joining rates for younger and less-educated teachers. This high joining rate for the young is also probably explained by the simple fact that, once a young person is trained as a teacher, a fairly logical first step in the job market is a job as a teacher. However, young teachers also *leave* in high numbers. One explanation might be that once they have been teaching for a year or two, they begin to understand what their lifetime prospects are and that their relative pay advantage is likely to decline; the more highly educated, the more they leave, because the more their lifetime earning prospects appear poor relative to the rest of the labour force. We can see that there are three groups who really represent very high leaving rates: the best-educated young and then the old in general. But note that amongst the young, the best-educated have higher leaving rates, and among the old, the least educated have the higher leaving rates.

One destination for the leavers is, perhaps, employment with school governing bodies. The following graph shows how neatly the ratio of SGB-employed teachers matches the rate of leaving from the PERSAL (public) database. Also note that the differences in the ratio of SGB to total teachers by age is statistically highly significant: middle-aged teachers are much less likely to be teaching in SGB posts than are the young and the old. Unfortunately, we did not have at our disposal similar data for independent school teachers.[18]



6. Forecasting Basic Numbers

In this section we attempt a little crystal-ball gazing to forecast the basic supply and demand for teachers.

It must be understood that forecasting something so susceptible to social trends and policy shifts as teacher supply and demand is extremely hazardous. Furthermore, these sorts of forecasts have absolutely no rigorous confidence intervals; that is, we cannot state our confidence in the results with any degree of precision, such as in saying that we are 95% certain that the forecast for the year *t* lies in the interval [*x,y*]. The demand side is relatively easy to forecast, but the supply side (and therefore the gap between supply and demand) is really quite chancy. All we can say, therefore, is that these are not true forecasts, but simply conditional projections; that is, projections conditional on a whole host of assumptions, some of which have to do with the likely course of, say, normal demography or the AIDS epidemic, and some of which have to do with possible policy choices in the future.

In this sense, we even make forecasts that we know with certainty cannot happen. We do that to be "counterfactual"—to analyse what would have happened, given government policy, if the assumptions government had fairly good reason to make in the early 1990s had worked out. This is an ex-post test of how "on-target" policy was, given the information then at hand. We will see that policy was actually fairly on target. Perhaps more analysis of the kind we are doing now should have been conducted, so that more information could have been available. But, given the information at hand at that point, the policy decisions appear reasonably rational.

To give the reader an idea of the accuracy of these sorts of forecasts: if a TV psychic is to be rated as having an accuracy of 0, and the physics of the projection of planetary orbits are rated at 10, our work is somewhere between 4 and 6. One might be consoled that 4 is better than a TV psychic, or one might be dismayed that 6 is so far from physics. Obviously, we find 4 consoling or else we would not bother writing this.

It should be noted that these sorts of projections are, in some ironic sense, perhaps more reliable to the degree we take them as "broad brush" and medium- to longer-term as opposed to short-term. For technical reasons beyond the scope of this report, such projections are actually particularly unreliable in the first few years. In the longer term, powerful determinants, such as basic demographics, whose parameters are more solidly known and whose roots are profoundly sociological rather than economic or political, tend to exert more influence than short-term phenomena such as teacher training or promotion policy.

One way to urge caution is to list the assumptions that need to be made in this kind of analysis, the relative impact each has on the forecasts, and the degree of faith we have in each one. Such a list should serve to highlight the fact that these sorts of projections are, in some sense, nothing but a highly systematic and transparent way of piling up assumptions, and these assumptions lead to certain conclusions because they

---

[18] Note that for 1998 we did not have as many data points as for 1997, so we display a fitted curve through the 1998 data points, whereas the line through the 1997 data points is not a fitted line.

are so systematically "piled up" and related to one another. (This is a crude way of describing the mathematics of this process, but it is apt.) To the degree that this "piling up" process is indeed highly systematic, and to the degree that the process forces one to make one's assumptions explicit, however, these sorts of forecasts can drive a dialogue process that can make the bases of a collective opinion clearer, and therefore create a legitimated choice for the way to proceed. But it is not physics.

Having made all these disclaimers, we must advise the reader that projections made by any analyst are subject to the same assumptions. Having to rely on these assumptions is not a weakness peculiar to our analysis. All modelling and forecasting is based on dozens of assumptions. Modelling and forecasting of social issues such as teacher supply and demand is particularly chancy because the assumptions are less certain and are themselves subject to policy change more so than is the case for physical phenomena. In this respect this exercise is no different from other exercises of this nature. We are perhaps being more candid with the reader than social and economic modellers and forecasters usually are.[19]

Because of the extraordinary uncertainty surrounding all these issues, only a tenderfoot analyst would put forth a projection as a line or as a set of points. Instead, below we present a reasonable lower bound on the gap between supply and demand, and a reasonable upper bound, with a range of scenarios in between.

The table below lists the assumptions and sets forth some discussion as to the range of values these take in the lower and higher bounds on the gap between supply and demand, as well as some observations on their historical rate.

Perhaps the greatest effect on the projected gap between supply and demand relates to the assumption that orphans require any special care, and the degree to which there needs to be a "special" L:E ratio, or simply orphan to care-giver ratio. We quite realise that in a system under heavy attack from the epidemic one may simply not be able to afford a desired L:E ratio even for non-orphans, and that therefore it may seem absurd to discuss scenarios with a special, lower, L:E ratio for orphans. However, we wish to portray the gap between the system's capacity to produce caregivers and the number of caregivers available as accurately as possible, because there are in fact policy choices that can be made to affect the number of caregivers available. These choices will have to be faced and will have to be made. The more accurately we can assess the urgency and importance of making these decisions, the more, we hope, we will contribute to ensuring that the decisions are wise, or at least well-informed. Thus, in this sense our modelling of this issue is strictly in the realm of "policy dialogue" rather than in the realm of "planning."

[19] This also explains why two totally different forecasts can both be "true": forecasts of this sort are really to be taken as true to the assumptions that drive them. Unlike in physics or even econometrics, these sorts of models cannot be true to reality. Therefore, two forecasts can be quite different from each other, depending on purpose, and both can be useful in understanding the *current* dynamics that are shaping up the future. If we had 10-50 times more and better data than we have, we could indeed see how well the model tracks reality, by starting the model in, say, 1990, and seeing how well it tracked, say, enrolment in the 1990s. However, the problem is that even in this case there were policy shifts in the 1990s that were exogenous to the modeling, and that would have affected the key parameters, in complex and endogenous ways.

Finally, note that in the table below some scenarios are totally "counterfactual." They will not happen; they cannot happen anymore. They are meant only to illustrate what might have happened if the assumptions under which the current policies were set in place had been correct. The point is to demonstrate that with the state of knowledge one had in the mid-1990s the current decisions were not quite as misguided as armchair critics would now suggest. (Though one would have to admit that not all the information on hand was used.) In any case, hindsight, as the cliché goes, is 20-20, and it is easy to be wise in retrospect.

Table 7. Assumptions needed to drive a teacher demand and supply projection			
Assumption	Importance	Degree of reliability	Discussion of numerical values
Picking out a particular demographic projection	High	Medium, and hides at least another 10-15 assumptions	Uncertainty exists as to whether size of population group of school entry age is already declining after having peaked in 1995 or so. The author believes it is, based on analysis of four data sources.
Picking out orphanhood scenarios	High	Medium	Whether orphans should "drive" teacher demand via a specific caregiver-to-orphan ratio, even if for dialogue purposes, and what that ratio should be, is a matter for discussion. Various orphanhood scenarios can be chosen, from a low of 24.3% of the cohort of 5-9 being orphans by 2015, to a high of 28.7%.
Intake rate into grade 1	Medium	Medium	This has been at least 1.0 for many years. It might conceivably decline. However, it is assumed to be 1.0.
Repetition rate in grade 1	Medium	Low	Repetition in the late 1980s and early 1990s was extremely high—as high as 35%-due to recycling of children admitted to school too young, under the assumption they would repeat. Controlled age of admission can reduce this apparent waste. It is assumed repetition declines to at most 10% (in a high enrolment scenario) or 5%) in a low enrolment scenario).
Grade-to-grade apparent or net flow ratios	High	Medium	These have been relatively low, historically, averaging about 93% (for an apparent loss of 7% between grades). In a high enrolment scenario it is assumed this goes up to 97% on average. In a low enrolment scenario it is assumed it stays at 93% on average. This is not to be confused with a true retention ratio.
Desired class size	High, but policy-driven, not a true assumption	High	This is not so much a matter of empirical analysis as a matter of policy setting, either real policy setting or idealised goals. It is assumed at 38 and 35 for primary and secondary respectively.
Period load ("work effort") of teachers; proportion periods taught (includes possibility that principals may have a very low teaching load)	Medium, though to some extent policy-driven rather than a true assumption	Low	This is partly a matter of empirical analysis and partly a matter of policy setting. It is assumed constant at 92% and 87% for primary and secondary respectively. This is a fairly good approximation of current reality.
Rate of substitute teacher usage	High, to some degree policy-driven, not a true assumption	Medium	Partly a matter of policy, partly a matter of necessity. It is assumed that in a best case scenario 2% of teachers at any moment need to be substituted. In a worst case it is assumed 5% will need to be substituted.
Usage of a "special" learner:educator ratio for orphans	High, to some degree policy driven	High	Can be set to any level. A 0-1 switch controls whether the ratio is used at all. The switch is 1 in the worst case scenario. If used, a 10 to 1 ratio is set, but can be re-set.
Assumptions about	Low for overall	High	Not explicitly used to drive overall conclusions. Assumed that as many as 25% would

Table 7. Assumptions needed to drive a teacher demand and supply projection			
Assumption	Importance	Degree of reliability	Discussion of numerical values
qualifications distribution of teachers	balance, high for skills distribution		remain under-qualified in primary level even by 2015.
Normal attrition rate of teachers	High	High accuracy of measurement in base year, but extremely susceptible to policy measures.	Measured to be 5.5% (depends a bit on qualifications), and staying at that level. Evidently this ratio is a key driver in the projections, and is very responsive to incentives and actual demand for teachers, as noted in the analysis of compensation and entry and exit rates in this document.
Mortality assumptions for teaching force based on AIDS scenarios	High	Medium	Normal attrition counts normal mortality. This refers to extra mortality from AIDS. Ranges from 3.5% by 2015 in the best case to 4.6% in the worst case.
Percent of grade 12 who sit for Senior Certificate Exam	Medium	High	This refers to the percentage of learners in grade 12 who opt for entering and sitting for the Senior Certificate Exam. This ratio is measurable with great accuracy at any given moment, but it is highly susceptible to policy and therefore its future variability is inherently difficult to forecast. A focus on increasing the pass rate could encourage schools to discourage learners who are unlikely to pass from sitting for the exam. This ratio has been at somewhere between 90% and 94% in recent years. It is assumed to stay fixed at 92% in any scenario.
Pass rate on Senior Certificate Exam	Low	High	This refers to the well-known “matric pass rate.” It is known with great accuracy in the base year, but it can fluctuate, as was obvious in the results of the year 2000. It is assumed to improve by 3 percentage points in the best scenario and by only 1 percentage point in the worst scenario. The base level is not assumed to be as high as the actual value for 2000, under the cautious supposition that 2000 may have been an exceptional year.
Ratio of headcount enrolment in TTC or tertiary institutions to Senior Certificate passes	High, extremely susceptible to policy	Low	This is a key driver in the system. It is not known with any accuracy whatsoever in the base year, and, on top of this, it is highly susceptible to policy shifts. This ratio drives the enrolment in teacher training institutions. It has been as high as 0.5 in recent memory, when the teacher training institutions were producing bumper crops of teachers, and has sunk to as low as about 0.11 (assuming our best estimates are correct) in 2001. In the best case scenario it is assumed this goes up by 5 percentage points each year. In the worst case it is assumed it decreases to 0.1 and then maintains that level.
Exit or graduation rate from TTC or tertiary institutions	High	Medium	This is a key driver as well. It can be known with reasonable certainty for the base period, based on its value in recent history. It measures, to some degree, the “internal efficiency” of the teacher training institutions, as well as the nominal length of the training programmes. (E.g., with a training programme of four years of nominal length,

Table 7. Assumptions needed to drive a teacher demand and supply projection			
Assumption	Importance	Degree of reliability	Discussion of numerical values
			this ratio would be 0.25.) The ratio is assumed to start at 0.25 and improve to 0.3 in the best case scenario, on the assumption that efficiency can be improved and/or programmes will be shortened. It is assumed to decrease to 0.2 if programmes are long and/or internal efficiency is low.
Percent of those enrolled in TTC or tertiary institutions who are already teachers	High	Medium	Ratio of enrolment that are already teachers and therefore cannot be assumed to drive a replenishing of supply. This has been at approximately 0.45 in recent history. It is intrinsically hard to measure this ratio because the necessary data have not typically been reported in any data documents in the past. Some recent measurements give us some degree of confidence. The ratio is assumed constant at 0.45. Not enough is known about the behaviour of this ratio to justify making any other assumption.
Percent who go on to pursue teaching upon exit from TTC or tertiary institution	High, extremely susceptible to policy	Low	This is a key driver in the system. Its historical value is not known with any certainty because it has never been measured. Furthermore, this ratio is extraordinarily susceptible to policy shifts. It is assumed to have a value of 0.5 in a historical base, to decrease to 0.4 in a worst case scenario, and to improve to 0.9 in a best case scenario.
Base year data on enrolment in all levels of the system	Medium (one might think this is high, but in fact, since in most projections what is interesting is the change over the base, this has only medium importance)	Medium for school level, low for tertiary institutions for 2001	Considerable uncertainty surrounds base year numbers. Discussion of these numbers would take us too far afield. As noted, we are more interested in changes over the base, rather than in absolute numbers per se.
Unit cost data for education programmes in TTCs and tertiary institutions and the effect of economies of scale on such costs	High	Low	This number is not known with much certainty because the cost of places in Colleges of Education in the last few years cannot be taken to be a reliable indicator of true cost, given that they have not operated at maximum efficiency. Costs at tertiary institutions in recent years are somewhat difficult to track because not all institutions report their data. Furthermore, the cost drivers are normally given in terms of Full-Time-Equivalents and other concepts because this is what the funding formula requires, but it is hard to link this up to the headcount concepts that are relevant for projecting teacher supply.

Using particular values of the assumptions listed above produces a set of scenarios for future teacher supply and demand. We present a set of scenarios below. Only the gap between demand and supply is shown. If the gap is positive, this means that demand is greater than supply: there is a shortage. If the gap is negative, it means supply is greater than demand: there is a surplus.

Table 8. Conditional forecasts of gap between demand and supply

Scenario	Explanation of scenario	Yearly gap 2001-2005	Yearly gap 2006-2010	Yearly gap 2011-2015	Demand (needed production) in mid-decade
1	Presumed policy intention, assumption of declining enrolment due to early decline of population due to fertility transition, tightening of age controls, etc., cutback in teacher training programmes, but no radical reaction by students choosing not to enrol in teacher training programmes. Assumed no AIDS impact known.	4000	-4000	-2000	11000
2	Same as scenario 1, but students react by radically choosing not to enrol, in reaction to lack of jobs due to relative hiring freezes in mid- and late-1990s; assumes that such a reaction is permanent. Assumed no AIDS impact known.	7000	3000	7000	11000
3	Assumes slower fertility transition (so no early population decline due to this factor), but worst-case AIDS scenario. No special attention to AIDS orphans via standard teacher training and supply. Assumes student over-reaction as in scenario 2.	13000	17000	21000	25000
4	Same as 3, but students' choice of teacher training as a post-secondary option goes back up to nearly traditional levels. Requires clear hiring messages from policy authorities, as well as low-cost study options for those choosing to be trained as teachers.	8000	2000	1000	25000
5	Similar to 3, but, on top, it is assumed AIDS orphans require special attention (L:E ratio of 10 to 1 for orphans). In addition it is assumed that student over-reaction is reduced by ½.	29000	38000	48000	57000
6	Similar to 5, but assumes that students' choice of teaching as a post-secondary training option goes back up to the level of the early 1990s, or five times (400%) the current level	28000	20000	12000	57000

Source: calculated by the author.

Scenario 1 shows that, if the assumptions presumably made had worked out, the policy reaction would have hit the mark. There would have been a small over-supply, in the medium term, in fact. But, considering that the employment base is some

350000, a mistake of a few thousand is not bad. Note that in any case there would have been a shortage in the early part of the decade. The assumption is that the cutback in enrolment would have come faster than any increases in the graduation rate from programmes, or in the proportion of students who actually join the teaching profession. Thus, there is a gap in the early part of the decade and then a surplus later on, but enrolment was gradually approaching a fairly good balance.

Scenario 2 shows what presumably happened as students reacted to poor hiring prospects due to hiring freezes and general teacher labour market prospects in the mid 1990s. In a sense, students have reacted to both the poor hiring prospects and the closings of teacher training colleges by assuming that job prospects will be forever as bleak as they were in the late 1990s. Traditionally (in early to mid 1990s), in South Africa, some 15% (very approximately) of matriculants chose to study teaching. By 2001 at best some 3–5% are choosing to study teaching. Even assuming an improvement in the graduation rate, and an improvement in the proportion of graduates who become teachers, there is a shortage starting now, which continue. However, the shortage is not as severe as one might think, because population growth is down, and enrolment is down because enrolment bloating in Grade 1 is under control.

Scenario 3 shows what happens if AIDS affects the country's teacher supply according to how one now fears might be the case, but students' low-enrolment reaction continues. Very large yearly deficits in supply show up very early. Cumulative yearly gaps as high as 20000 show up by 2010.

In scenario 4 we assume that eventually students themselves adapt to the fact that hiring will not be permanently frozen. It is assumed that authorities send clear messages that hiring will resume (depending on authorities' cognisance of the need to re-open hiring due to AIDS deaths). Possible changes in policy also shorten time of training or ease entry requirements. Eventually the shortage is addressed. But note that it might take some time (depending on how clearly the message is sent to students) to change student behavioural choices. If the messages are clear and strong, then the shortage can be addressed more quickly, instead of having to wait until 2006 or so. The scenario assumes only a very slow reaction by students.

Scenario 5 assumes that AIDS orphans require special attention, and assumes that student over-reaction is reduced by ½. That is, it is assumed that student choice of teaching as a post-secondary option is improved by 100%. Even then, a yearly shortage of some 38000 and then 48000 teachers develops.

Scenario 6 is the same as scenario 5, but assumes an easing of student perceptions and a growth in teaching as a choice, of 400% (five times current levels). Even then there is a yearly shortage of 12000. This scenario is fairly important. It suggests that even in the extremely unlikely scenario that students go back to choosing teaching in the same proportions they did in the early 1990s, there would still be a very large shortage in the sector's ability to use standard teacher training and/or allocation as a means of dealing with orphanhood. Note that the shortage is actually most acute in the near term, because it is assumed that it would take time to change students' perceptions. Clearly, more informal or community-based means of addressing orphans' needs will

have to be found. This important point ties in to the next set of issues, discussed below.

In conclusion to this section, we believe the following cautious statements are justifiable.

- If the state attempts to pay special attention to the orphanhood generated by AIDS, and attempts to do so with formally trained teachers, then we simply do not see any scenarios that make this feasible, unless society's resources were to be really skewed towards this one task, and most other educational tasks dropped or totally minimised.

- If the state decides not to pay special attention to orphans, or allows that it simply cannot do so with formally trained and formally paid teachers, then it is possible to think of scenarios wherein the system could train sufficient numbers of teachers to cope with AIDS amongst teachers, and then address the issue of orphanhood through more informal and community-based means. Essentially, "all" that would be required is a return to the basic social parameters of transition between Grade 12 and teacher training that were common in the early 1990s. One would have to plan so that some 15% or so of matriculants would choose to become teachers, instead of the 2%–3% (apparently—there is uncertainty) that are making this choice now. Our research suggests that this requires not so much a shift in pay policy or salary scales, as simply a) for the state to start sending out clear informational messages to secondary school students, about the likelihood of jobs being available in teaching (obviously these messages should ideally be based on further research to confirm whether we are right in this paper), and b) improving the planning and enrolment capacity of cost-effective trainers of teachers.

- The latter would require considerable analysis of options beyond the scope of this paper, for the cost-effective training of teachers. However, it does seem that one could conclude that teachers should not be trained in the relatively expansive, and perhaps relatively inefficient, method common in the early 1990s and certainly common in the late 1990s. We calculate, for example, that the social cash cost (i.e., public and private, but without counting the private opportunity cost) of one year of pre-service training, by the late 1990s, was approximately R45000 to R50000 per student per year in teacher training colleges, and approximately R20000 per student per year at higher education institutions. However, note that this is probably not a "fair" comparison. It is likely that by the late 1990s, teacher training colleges were running well below optimal scale. At a time when they were functioning at levels closer to optimal scale, say in the early 1990s, the cost in today's Rand would have been about R20000 per student, or about the same as at a tertiary institution.

- However, the issue of the real effectiveness of the training would have to be faced. While some South African research suggests that pre-service teacher training is an effective form of expenditure for the state, it is unclear whether it is the subject-matter content of the training, or other aspects of the training (including mere selection, or teacher training as a filtering device: some colleges simply acting as a filter for students who would be better teachers than others no matter where they study), that has made pre-service teacher training an effective intervention. Three

to four years of pre-service training, then, would cost somewhere around R60000 R100000 per teacher. Scenario 4 above, which is what is needed to reach some form of balance, suggests training some 30000 new teachers per year. At a cost of R80000 (to take the mid-range) per teacher, the price tag is about R2.5 to R3 billion per year. One ought to make sure such an expenditure is cost-effective. Thus, before embarking on the sorts of scale-up suggested in the paragraph above and in Scenario 4 as a way of facing AIDS, we would suggest that this matter be addressed as a matter of utmost urgency.

7. Regionality and Micro-Regionality of AIDS Epidemic

In this section we want to make a simple point: that the AIDS epidemic strikes in a highly selective manner. This implies that the administrative measures used to confront the epidemic will need to be localised or, alternatively, much more directive.

The following table shows HIV prevalence rates at 30 KwaZulu Natal hospitals in 1998. The hospitals' names have been omitted to protect anonymity and to draw attention to the numbers themselves rather than to the hospitals. A certain degree of homogeneity has been ensured by focusing on provincially managed hospitals as such, excluding clinics, private hospitals, etc.

Table 9. HIV prevalence rates at KwaZulu Natal Hospitals, 1998	
Case 1	17.0
Case 2	18.0
Case 3	21.3
Case 4	21.3
Case 5	22.0
Case 6	24.0
Case 7	25.0
Case 8	26.7
Case 9	27.0
Case 10	27.7
Case 11	28.0
Case 12	28.4
Case 13	28.5
Case 14	29.4
Case 15	31.0
Case 16	31.0
Case 17	31.6
Case 18	32.0
Case 19	33.0
Case 20	33.0
Case 21	33.5
Case 22	33.6
Case 23	34.0
Case 24	36.0
Case 25	37.7
Case 26	45.7
Case 27	45.9
Case 28	46.0
Case 29	50.0
Case 30	58.0

Source: personal communication, Daniel Wilson, EduAction, Durban. Original Data: Medical Research Council, Department of Health.

These prevalence rates may or may not reflect likely prevalence rates amongst teachers. However, the age groups at which these prevalence rates apply, and the gender to which they apply, are actually fairly coincident with the demographics of the teaching force. But let us be conservative, let us use these numbers to help us reason in terms of schools, and let us assume that the variability (not necessarily the mean) HIV-prevalence rates at the hospitals bear some reasonable relation to the variability in eventual AIDS-incidence rates in regions but a little lower. It seems safe to assume that in some schools 1 teacher out of 10 will be affected by AIDS, and will likely die, whereas in other schools, 4 out of 10 teachers will be affected and will likely die.[20]

It is unreasonable to suppose that any system of human resources allocation that is centralised (in actual resource allocation using a provincially driven post allocation

[20] Note that this does not mean that the death rates at those schools in any given year will be 10% and 40% respectively. All other things being equal, if the disease lasts, say, seven years from infection to death, a prevalence rate of 50% would be approximately equivalent to a death rate of 7% (50/7). To see why, note that the prevalence of death amongst humans is 100% (we will all die sooner or later), but it takes us about 65 years to die, on average, so the death rate at any moment is about 100/65 or 1.5%.

model) and yet participatory (in consulting many stakeholders for every transaction) will be able to cope with this problem. The problems of *participatory* coordination, but from a *centralised* resource allocation will simply overwhelm the administrative capacity of the system. It would seem that the human resources provisioning system will have to be either far more directive and non-participatory (the locus of allocation matching the locus of deployment decision-making, and transactions costs being lowered) than it is now, ***or*** far more decentralised (again, the locus of allocation now being made coincident with the locus of deployment decisions). The present mix is quite attractive in some ways, but it can work only in a highly stable and predictable environment where its high transaction costs can be ignored.

One may be tempted to take comfort in the notion that these wide micro-regional or regional differences are to be expected only in a country in the early stages of the epidemic. Data from countries with mature epidemics, however, suggest (only suggest) that the problem gets, if anything, worse as the epidemic matures. The following data (table 10) from Uganda illustrate the point.

Table 10. Regionality of HIV prevalence in a mature epidemic, Uganda 1998

	Prevalence rate
Two lowest regions	
Matany	1.3
Pallisa	2.6
Two middle regions	
Mbale	6.3
Soroti	7.7
Two highest regions	
Kagadi	11.5
Mbara	10.9

Source: UNAIDS/WHO Epidemiological Fact Sheet, 2000 Update, http://www.unaids.org/hivaidsinfo/statistics/june00/fact_sheets/pdfs/uganda.pdf

As can be noted, the ratio of highest to lowest prevalence rates is about 5 to 1, thus, in fact, higher than that in South Africa. Countries other than Uganda show exactly the same pattern, but we are not showing the data so as not to clutter the presentation. However, note that the absolute magnitude of the prevalence rates is lower in Uganda than in KwaZulu Natal (admittedly perhaps the worst-affected overall region in South Africa). The administrative nightmare of dealing with a high variance really only shows up if the averages are also high, which is the case in KwaZulu Natal and in other badly affected regions of South Africa.

Annex A

Table A.1. "Multivariate" analysis of simple demographic trends, no sampling weights				
Equations where the dependent variable is either a quantity (e.g., salary) or the conditional probability of being in a group				
Dependent variable	(1) Dummy variable for being a teacher	(2) Simple time trend	(3) Interaction of time trend and teacher dummy variable	(4) Simple time trend for teachers only
Salary	+,***	+,***	+,***	Irrelevant
Salary, controlling for hours worked	+,***	+,***	+,***	Irrelevant
Probability of being African	+,***	+,***	-,**	+,**
Probability of being coloured	-,***	+,*	NS	NS
Probability of being Indian	NS	-,***	NS	-,***
Probability of being white	-,***	-,***	+,***	-,*
Probability of being female	+,***	+,***	NS	+,**
Probability of being a union member	+,***	+,*	+,***	+,***
Age	-,***	-,***	+,***	+,***
Hours worked	-,***	+,***	-,***	-,***
Equations where the dependent variable is the conditional probability of being a teacher, as determined by a particular variable				
Condition or group	Dummy for the condition or group	Simple time trend	Interaction of time trend and dummy for the condition or group	
Union membership	+,***	-,***	+,**	
Female	+,***	-,***	-,*	
African	+,***	-,*	-,***	
Coloured	-,***	-,** *	NS	
Indian	NS	-,***	NS	
White	-,***	-,***	+,***	

Source: OHS 1995, 1997 and 1999. Author's tabulation.

Annex B. Basic Income Equations

The following table shows the basic income determining equations used in the graphical analysis presented in the main part of the paper. Note that in the work represented by the following equations we were not interested in an in-depth understanding of the "ultimate" determinants of teacher vs. non-teacher pay. Instead we wanted simply to see whether there seem to be any systematic, interactive, and non-linear differences between teacher pay and non-teacher pay, depending on factors such as age and education, in particular. To some degree this statement is a bit disingenuous, in that evidently there has to be some notion of a causal theory behind any such attempt. However, it is nonetheless important to point out that we are not interested in deep causal analysis, because otherwise the exercise below would tend to appear excessively like a fishing expedition.

The OHS data, particularly when it comes to the income variables, are not as clean and reliable as would be ideal (see annex explaining clean-ups we had to carry out). We thus trimmed away observations whose residuals were more than 2.5 standard deviations above or below the mean residual. Obviously, this exaggerates the R^2. The values for the untrimmed and trimmed versions are shown. The t-values would also be exaggerated, naturally, but their relative sizes would nonetheless be of interest, so they are presented. There is no *a priori* reason to suppose that the value of the coefficients would have changed radically via the trimming process. While it is true that the trimming process naturally exaggerates the goodness-of-fit and the t-values, note that in the untrimmed versions of the same equations, the R^2 were actually quite high by the standards of these sorts of cross-sectional analyses and do not increase by a huge amount in the trimmed version—similarly with the t-values. The coefficients and t-values for the variables Female and $AfEd^2$ were -663 and 5.7, -36 and -9.1, respectively, in the untrimmed version, and -530 and 4.9, and -39 and -8.5 in the trimmed version. Thus, our substantive conclusions are not much affected by the trimming, and the trimming affords some protection against dirty data and outliers with undue influence.

Table B.1. Income determination equations

		1995		1997		1999	
Variable	Variable Explanation	Coefficient	t-value	Coefficient	t-value	Coefficient	t-value
(Constant)	Constant	4190.9	7.0	1210.0	3.8	2943.1	4.9
Female	Dummy for being female	-590.1	-39.4	-530.5	-36.0	-557.7	-28.6
Age	Chronological age	89.4	15.2	59.0	10.3	65.7	7.9
Age^2	Chronological age squared	-1.11	-18.2	-0.74	-12.9	-0.89	-10.5
Union	Dummy for belonging to a union	374.0	24.0	397.6	25.0	600.9	28.5
TotEd	Total years of education (approximate)	-951.4	-9.6	-235.8	-4.5	-724.5	-7.8
$TotEd^2$	Total years of education squared	54.2	13.2	21.5	9.6	46.6	12.6
African	Dummy for being African	-4972.5	-8.5	-1334.6	-4.6	-3071.4	-5.4
Coloured	Dummy for being Coloured	-5029.0	-8.5	-1410.8	-4.8	-3111.4	-5.5
Indian	Dummy for being Indian	-3629.2	-5.6	-500.4	-1.1	-2322.9	-3.5
Teacher	Dummy for being a teacher	-853.0	-1.5	-1074.0	-1.8	249.5	0.2
$TeachEd^2$	Squared interaction term for being a teacher and the square of years of education (square of TotEd)	-1.2	-0.6	-0.2	-0.1	-9.7	-2.9
AfEd	Interaction of years of education and being African	701.7	7.1	18.2	0.4	372.1	4.1
CoEd	Interaction of years of education and being Coloured	748.5	7.6	19.5	0.4	403.0	4.4
InEd	Interaction of years of education and being Indian	446.6	4.0	-66.3	-0.8	255.2	2.3
$AfEd^2$	Interaction of years of education, squared, and being African	-35.0	-8.5	-4.9	-2.3	-23.2	-6.4
$CoEd^2$	Interaction of years of education, squared, and being Coloured	-35.7	-8.5	-1.1	-0.5	-20.5	-5.4
$InEd^2$	Interaction of years of education, squared, and being Indian	-16.0	-3.2	1.4	0.4	-11.7	-2.3
EdAge	Interaction between years of education and age	3.04	6.0	3.10	6.2	5.19	7.5
$EdAge^2$	Interaction between years of education and age, whole term squared	0.0003	0.6	-0.0002	-0.6	-0.0012	-2.1
TAge	Interaction between being a teacher and age	98.5	3.2	69.7	2.2	-43.4	-0.6
$TAge^2$	Interaction between being a teacher and age, squared	-0.92	-2.7	-0.96	-2.7	-0.28	-0.4
TEdAge	Interaction between being a teacher, years of education, and age	-4.05	-2.3	-1.67	-1.1	6.13	1.5
$TEdAge^2$	Interaction between being a teacher, years of education, and age, whole term squared	0.00	1.3	0.00	1.5	0.00	-0.9
TAf	Interaction between being a teacher and being African	657.9	8.3	807.6	7.4	1508.0	10.4
TCo	Interaction between being a teacher and being Coloured	984.9	8.8	1150.7	8.4	1429.2	7.1
Tin	Interaction between being a teacher and being Indian	-0.7	0.0	774.0	3.7	683.7	1.9
R2 and F stat, no residual trim		0.48, 872		0.39, 561		0.37, 426	
R2 and F stat, 2.5 standard deviations residual trim		0.59, 1321		0.54, 993		0.55, 880	

Source: calculated by the author.

Annex C. Record of modifications to OHR databases needed to carry out analyses

To create unique id one needs to concatenate with the concat(rtrim(ltrim(eanumber),...))) approach. Note that eanumber and so forth had to be turned into strings first.

Command is:
concat(ltrim(rtrim(mdnumber)),ltrim(rtrim(eanumber)),ltrim(rtrim(vpnumber)),ltrim(rtrim(ppersno)))

Then one needs to convert the resulting variable into a numeric one:
number(persuniq,F13.0)

mdnumber string 3
eanumber string 4
vpnumber string 2

ppersno string 2
uniquep string 11
uniquenu numeric 11
uniquehh string 9
uniquehn numeric 9

In worker file, educat = 99 was changed to educat = missing.

Education levels in the worker files in the variable educat were wrong, in my opinion, or at any rate not clear. Using the descriptions given by StatsSA, the following was done:

Values for phghqual = 9 were recoded as missing
Values for phghqual = 8 were recoded as missing
Values for phighshc= 99 were recoded as missing

This was done in both “workers” and “workers with income” files.

(Note that a file “workers with income” was created for workers whose income reporting field was non-empty.)

(In 1999 the "other" or "don't know" codes were 21 and 22 and these were recoded as missing in the “all worker” income files.)

Based was recoded as based = phighsch - 1 except for phighsch=0.

Then, toted was coded as:

toted = based + PHGHQUAL if PHGHQUAL is not missing
otherwise = based

A summary variable for education was created, called sumed, as follows, depending on the value of toted:

0 = 0
1-11 = 1
12 = 2
13,14 = 3
15 = 4
>15 = 5

Got rid of extremely high or odd values of salary and wsalamt: 1000000, 999999 , 9999, and 999.

Same was done for wservpr, wservsa1, wservssa2, wspgds, wspsal, and wspoth. Got rid of 999999. Declared as missing values.

In 1999 the relevant variables are:
Income of employee (Q3_20AEM),
Time period of payment (Q3_20BEM)
Income bracket (Q3_20CEM)
Income of self-employed (Q3_26AEM)
Time period of payment (Q3_26BSE)
Income bracket (Q3_26CSE)

For whowpaid and wservpai, zeroes were also declared missing values. Also for wservpr, but only in calculations in xls sheet for imputed income.

Created a variable called totcost which sums the costs for self-employed: wspgds+wspsal+wspoth=totcost. This was possible in 1997 but not in 1999, because in the latter year the costs of doing business for the self-employed were not reported.

Created a dummy variable that picks out cases where totcost is unreported (missing) but wserva1 or wservpai are non-missing as follows: (missing(totcost) & (~missing(wservsa1) | ~missing(wservpai))) | (~missing(totcost) & (missing(wservsa1) & missing(wservpai))). The dummy was called selfcons (=1 for inconsistent, 0 for consistent).

A file was created from workers for everyone for whom wsalamnt or whowpaid was not missing, in fact if (~MISSING(wsalamt) | ~MISSING(whowpaid)|~MISSING(wservsa1) | ~MISSING(wservpai)) & selfcons ~= 1.

For 99 it was ~MISSING(q3_20aem) | ~MISSING(q3_20cem)| ~MISSING(q3_26aem) | ~MISSING(q3_26cse).

This is a file for all workers who have declared some form of income, even if 0.

This file was called "workers with sal or self income." (Same for 99.)

An intermediate file was created to transfer just these variables to Excel: the uniquenu id, and wsalamnt, wsalave, whowpaid, wservpr, wservsa1, wservsa2, wservpai, wspgds, wspsal, wspoth (in 1997, in 1999 the cost variables were not used).
New values were created in Excel, in particular for people with actual salary amounts, these were all translated to monthly amounts. (Using the wsalave variable.) This was called salmonth.

A variable called imputed was used to denote those for whom imputation was needed as below.

Salaries were also imputed for those who only indicated a salary range. The median salary, for each range class, of those for whom there was both range and actual salary information, was used.

The basic analytical table used (for 99; for 97 we did not bother keeping the syntax to save on the details) was:

```
* Basic Tables.
TABLES
  /FORMAT BLANK MISSING('.')
  /OBSERVATION q3_20aem
  /TABLES q3_20bem > (q3_20cem) > q3_20aem
  BY (STATISTICS)
  /STATISTICS
  count( ( F5.0 ))
  mean( )
  median( )
  minimum( )
  maximum( ).
```

The result was analysed in an Excel sheet.

This was also done for the self-employed. Here we took care to adjust for months worked. For the self-employed, the median and mean income reported often did not jibe with the income categories, for people that reported income and category. To impute income, we used the actual average even when it did not agree with the income category.

For people claiming that the reporting period was yearly the actual amounts just did not make sense. So we did not take this into account when using ranges to impute.

Note that in the Excel sheet used for imputing values there were some non-visible characters in some observations for some of these key variables. They were not zero, nor totally blank, and were not visible. We blanked out some of these fields in constructing the following key created variables:

IMPUTED SALMON SELFMON SALIMP SELFIMP
TOTCOSTNUM TOTNETINC

Row 1 Row 2

A UNIQUENU 102311
B WSALAMT
C WSALAVE
D WHOWPAID
E WSERVPR 12
F WSERVSA1
G WSERVSA2 2
H TOTCOST 33500
I WSERVPAI 14
J IMPUTED IF(AND(B2=" ",F2=" "),1,0)
K SALMON IF(B2<>" ",IF(C2=3,B2,IF(C2=2,B2*4.33,B2*21.667))," ")
L SELFMON IF(AND(F2<>" ",E2>0),IF(G2=2,F2,IF(G2=1,F2*4.33,F2*12/E2))," ")
M SALIMP IF(B2<>" ",K2,IF(OR(F2<>" ",I2<>" ")," ",VLOOKUP(D2,U$15:V$28,2,FALSE)))
N SELFIMP IF(F2<>" ",L2,IF(OR(C2<>" ",D2<>" ")," ",VLOOKUP(I2,X$16:Y$31,2,FALSE)))
O TOTCOSTNUM IF(H2<>" ",H2," ")
P TOTNETINC IF(AND(M2<>" ",N2<>" ",O2<>" "),M2+N2-O2,IF(AND(M2<>" ",OR(N2=" ",O2=" ")),M2,IF(AND(N2<>" ",O2<>" "),N2-O2," ")))

Note that the adjustment for yearly income values for self-employed workers reporting less than 12 months was done only if the workers reported that their Rand income report was annual.

A dummy for teachers, teach, was created via the occupation codes 2310 through 2390 and also 3310 to 3330.

Dummies for four races were created.

Log and square versions of salimp, toted, and page were created.

Also versions of all these variables times the teacher dummy.

Unique person 10811513 was actually a duplicate in 1997. Removed.

In worker file the following additional changes were made.

Gender was transformed from 1=male to 0=male and 2=female to 1=female, and the variable name was changed to female. It was felt this was a more standard treatment.

Trade union was changed from 2=no to 0=no and 1=yes to 1=yes, and the variable name was changed to union; 0 was changed to missing. In 1999 3 was "don't know" and was changed to missing.

Hoursworked (whourw) was changed because 999 must be missing values. There is no other logical conclusion.

The imputed salary per normal month was calculated as salimp*40/hours usually worked.

salhrs - salary adjusting for hours worked previous week
salhrsu - salary adjusting for hours worked, usual week
totnethu - total net income adjusting for hours worked, usual week
totnethr - total net income adjusting for hours worked previous week
selfhr - self income adjusting for hours worked previous week
selfhru - self income adjusting for hours worked, usual week

To use the weights, the person weights were divided by the mean of perweight in 1997 or the appropriate weight variable in 1997.

Number of workers with data in various files

1995: 32044
1997: 25774
1999: 25428

Total 83246

Important. We discovered that the 1999 income data had a huge variance. We decided to filter out all values > 40000 in all analyses involving income.

Also, we filtered out observations which had standardised residuals > abs (2.5 or -2.5) in the main income determining equations.

In PERSAL, to create attrition rate, first sort by PERSAL number and then by year in descending order.

Stayed=0
Stayed=1 if year=98, and lagyr=99 and lagid=persalno

Sort by persal number and then by year in descending order

Gone=0
Gone=1 if (year=98 & lag(year)=98) | (year=98 and lag(year)=99 and persalno ~= lag(persalno))

New:

(year=99 & lag(year)=99) | (year=99 and lag(year)=98 and persalno ~= lag(persalno))