2001

Developing schools' capacity to make performance judgements

William Loudy

Helen Wildy

DEVELOPING SCHOOLS' CAPACITY TO MAKE PERFORMANCE JUDGEMENTS

Final Report
December 2001

William Louden & Helen Wildy

Institute for the Service Professions

Edith Cowan University

Western Australia

Contact:
w.louden@ecu.edu.au
http://isp.ecu.edu.au/dataclub

TABLE OF CONTENTS

INTRODUCTION

The purpose of the grant was to identify and document best practice in supporting school communities to make judgements about performance in relation to benchmark standards.

The initial proposal for this project specified a 12-month time-line, commencing February 2000 and being completed by November 2000. The early success of the project lead to reconsideration by the Education Department of WA of the scope and duration of the work. A contract variation submitted to the Department of Education Science and Training in May 2001 led to an extension of the project until November 2001.

An independent evaluation of the project has been prepared by AAAJ Consulting, who contributed a chapter on the links between policy, practice and research in the report *The Impact of Educational Research* (2000). The evaluation (included as Appendix 1 of this final report), drew the following conclusions:

> Overall, we were impressed by how open the principals were to learning more about the data coming from their schools. Before the Data Club it was not at all evident that the great majority of primary school principals would take on board the results of national tests in examining the school's practices and strategies. The very favourable response to the offer of help in analysing and understanding the WALNA data by the principals is an important finding in its own right (p. 8)
>
> The information gathered from this survey about the *use* being made of the Data Club by principals (and incidentally by Districts) suggests that the data, the analysis, and the skill and understanding developed through participation in the Data Club is being applied extensively.… In the process of working through the Data Club, principals (and schools) appear to have stopped being defensive if they are not performing as well as expected (p. 14).
>
> In sum, the professional development aspect of the Data Club has provided much needed tutoring about analysing and using the WALNA data in schools, and it has been very much appreciated. There was not a single principal who felt that he or she had not learned what was intended for them to learn. The outcome from such successful PD is that principals want more – more for themselves and more for their teachers (p. 16).
>
> The Data Club has begun very well, but its role has only just begun. Schools recognise that there will be much more for them to learn about using data over the next few years. And they will want reliable help from independent experts. The Data Club has provided those services to everyone's satisfaction – indeed it seems to have exceeded expectations (p. 18).

The project brief identified four tasks, which are listed below. An account of progress against these tasks follows.

1 Literature review comprising review of Australian and international literature documenting best practice in provision of comparative student performance data to schools.

2 Options paper comprising set of options for representing State-wide benchmark assessment data to schools.

3 Data display and self-evaluation strategies to assist school communities to make fair, accurate and defensible inferences about school and student performance.

4 Undertake consultations with schools, school district staff, parents and community groups and trial a bureau service that:

- explains the data displays, analyses benchmark data and self-evaluation strategies;
- enables schools to undertake self-evaluation using benchmark data;
- identifies strong and weak inferences that maybe made from the data;
- builds schools capacity to explore and use their own and other comparable data for self improvement and school accountability purposes.

PROGRESS AGAINST TASKS

Task 1

Literature review comprising review of Australian and international literature documenting best practice in provision of comparative student performance data to schools.

Progress

The literature review was provided to the Department in September 2001 (see Appendix 2). It provided a background analysis concerning the relationship between student performance data and school improvement, and technical discussions of the 'like-schools' and 'value-added' analytical strategies. Five issues were identified:

- What grounds are there for choosing between like-schools and value-added analyses?
- How should socio-economic status be accounted for in the analysis?
- What advantages are there in the use of group or individual assessment data?
- How much precision is possible, given the characteristics of the data set?
- How should statistical uncertainty be represented in data displays for use by schools?

On the basis of this analysis, seven recommendations were made for the use of the Western Australian Literacy and Numeracy Assessment (WALNA) data:

1. In the absence of adequate time-series data from the Western Australian Literacy and Numeracy Assessment (WALNA), schools make use of like-schools analyses.
2. As adequate time-series data become available from WALNA, schools make use of value-added analyses.
3. The Ross Farish H index be the area-based measure of SES used in WA like-schools calculations.
4. The adequacy of 'H' in comparison with individual measures of SES such as parental education and income be investigated in a sample of WA schools.
5. Consideration be given to the introduction of unique student identifying numbers, to improve the precision of like-school and value-added calculations.
6. In the absence of unique student identifying numbers, simple regression rather than multilevel modelling be used to make like-school and value-added calculations.
7. Statistical uncertainty as well as calculated school values be represented in like-school and value-added data displays.

Task 2

Options paper comprising set of options for representing State-wide benchmark assessment data to schools.

Progress

The options paper, which built directly on the analysis in the literature review, was provided to the Department in September 2001 (see Appendix 2). Three data display strategies were identified and documented in the options paper:

- student performance box-and-whiskers plots,
- like-schools residuals plots, and
- value-added residuals plots.

Task 3

Data display and self-evaluation strategies to assist school communities to make fair, accurate and defensible inferences about school and student performance.

Progress

The data display and self-evaluation strategies developed by the Data Club and shared with schools through 70 face-to-face workshops are available on the Data Club web-site <http://isp.ecu.edu.au/dataclub>. Appendix 3 provides:

- a description of the feedback provided to schools;
- worked examples and advice on interpreting the feedback;

- a downloadable, web-based graphing program; and
- professional development resources including a downloadable PowerPoint presentation and order details of videos of the Stage 1 and Stage 2 workshops.

Task 4

Undertake consultations with schools, school district staff, parents and community groups and trial a bureau service that:

- explains the data displays, analyses benchmark data and self-evaluation strategies;
- enables schools to undertake self-evaluation using benchmark data;
- identifies strong and weak inferences that maybe made from the data; and
- builds schools capacity to explore and use their own and other comparable data for self improvement and school accountability purposes.

Progress

The initial project proposal was based on a pilot group of 150 schools. Demand from schools led to 459 schools participating in 2000 and 506 schools participating in 2001 (see Table 1). The participation rate (number of Data Club schools as a percentage of total Government schools with year 3 or 5 students) has grown from 55% in the first round of Stage 1, an additional 15% in the Stage 1 catch-up workshops, and a further 7% in Stage 2. The final 2001 participation rate was 77%, involving schools from all 15 State Education Districts. A complete list of schools participating and workshops offered appears as Appendix 4.

Table 1. Data Club Participation, 2000-2001

	Workshops	Districts	Schools	Participation rate
Stage 1	27	14	362	55%
Stage 1 Catch-up	10	10	97	15%
Total Stage 1	37	13	459	70%
Stage 2	33	15	506	77%

Appendix 1: The Data Club: Its Impact on WA Primary Schools

The Data Club: its impact on WA primary schools

Report prepared by AAAJ Consulting Group
Jane Figgis & Anne Butorac
23 November 2001

Thirty primary school principals were interviewed about their participation in the Data Club. The set of open-ended questions probed why they had become involved, whether (and how) the Data Club was proving useful, and exactly how the national testing data (referred to in Western Australia as the Western Australian Literacy and Numeracy Assessment, WALNA) are now being used. The phone interviews ran from 10 to 30 minutes, with most at the longer end of the spectrum. The random sample of 30 schools (from the total 505 schools participating in the Data Club) was generated using SPSS's automatic random selection process. Phone interviews about the Data Club were also conducted with four District Offices (chosen arbitrarily), and six individuals nominated by Dr Helen Wildy.

The most immediate and striking finding from these conversations is that prior to the Data Club very few schools were using the WALNA data. Corroborating evidence comes from the Australian Council for Educational Research, which conducted the testing in 1998. Schools were asked to bring their 1998 data to the Data Club professional development session - 450 schools out of the 505 couldn't find the data and each had to pay ACER $27.50 to get a replacement copy.

This report analysing the interviews is in four sections:

1 Why principals signed up for the Data Club;
2 The use now being made of the WALNA data – in schools, at District level (and elsewhere), the particular use of the website;
3 The professional development provided as part of the Data Club;
4 Related issues (including confidence in WALNA testing).

Each section concludes with a summary highlighting the implications of the findings.

1.0 WHY PRINCIPALS JOINED THE DATA CLUB

The most frequently cited reason for joining the Data Club was the opportunity to compare the school's results with *like* schools. This was the first opportunity principals have had to make that comparison and they found it 'satisfying' (the word they often used) to be able to do so. It wasn't that they were always satisfied with the result, but a critical yet unanswered question was now being answered – a mystery solved – and that felt good:

> *Every school works in isolation – what we think is super duper may not be that at all.*

> *I'm interested in being able to compare with other schools – especially other schools in the same band.*

The difficulty about testing for us has always been that we compared with state norms and we always looked terrible. The end result was always miles below, even though teachers have worked hard. Very hard for people to feel good about themselves … And what we hoped would happen has. We can really celebrate some of our achievements because we're now comparing the old 'apples to apples' – that has helped enormously. We know that at the end of the day we also have to compare ourselves with every Year 5 kid in the state, but knowing we're doing it compared to where our kids should be at helps.

Before the Data Club we would have a squiz at the data, and had a bit of an idea about how we were going, but we didn't talk to other schools about it and certainly didn't have a good picture of how we compare with other schools. That's what I was hoping for – that comparison with like schools.

Most principals were keen to use the WALNA data – a sentiment that has perhaps not been fully appreciated before. That they had not been doing so was primarily because they did not know exactly what the data meant. In fact, to skip ahead a bit, what they so consistently got from the professional development sessions was the ability to *see* - quite literally - the educational information in the data. That was a revelation for the majority of principals. The reasons principals gave for their initial interest in the Data Club included:

We get all this information and we don't really know what we should do with it or how to read it. The Data Club gave us a chance to have a look at some serious analysis of the data, and get some expert advice.

I'm a new principal and I wanted to understand this data. I hadn't been too familiar with it before. I hadn't been involved in data analysis.

There was a lot of opposition in the school – a feeling that the tests were not a true reflection of the students - so I needed to understand it. This kind of accountability isn't going to go away.

The appeal of getting some assistance – feeling that any help would be useful. It's not so much that we were struggling with the WALNA data before (we could analyse it) but not to the same extent as now.

For me it was a matter of actually being able to have some PD – so I could gain a better understanding.

It provided better stats than we could produce ourselves.

We need to be able to talk to the District Directors about the school data, so I had to learn how to do that.

I'm interested in school improvement – and you need good information for that.

One of the other significant attractions of the Data Club mentioned by the principals is the capacity it will give them in the future to track student cohorts over time.

A District Director pointed out that one of the simple things the Data Club did – quite apart from the data analysis - was to inject space and time into principals'

busy lives. Without the Data Club many of these principals would simply not have made the concentrated time available to think carefully about the test data.

There are schools where industrial action against national testing remains high on the agenda and who refuse to participate in the Data Club. They were not included in this survey, although it is worth pointing out that there are schools which have joined the Data Club - several of whom turned up in our sample - which still have concerns about the testing regime. The issue is discussed later in the report.

Overall, we were impressed by how open the principals were to learning more about the data coming from their schools. Before the Data Club it was not at all evident that the great majority of primary school principals would take on board the results of national tests in examining the school's practices and strategies. The very favourable response to the offer of help in analysing and understanding the WALNA data by the principals is an important finding in its own right.

2.0 THE USE NOW BEING MADE OF *WALNA* DATA

2.1 In schools

As a consequence of schools' involvement in the Data Club, much more use has been made of the data emerging from the WALNA tests. The use made of the data, and more particularly of the data analysis facilitated through the Data Club varied across the participating schools. However, a number of common themes emerged though this review.

Perhaps the most important observation is that, with one exception, the principals we talked to are looking their school data squarely in the eye and accepting that it has something important and relevant to say to them about their school. Where the results have been less than they would have liked, they have said so openly and honestly. Interestingly, even where the results have looked 'okay' in that the students were performing 'as expected' for schools in their 'H band', principals have dug down almost searching to find things that would disappoint them – and often succeeded.

In all except two of the schools we spoke with, the principals have discussed the Data Club approach to analysing WALNA data with the whole of school staff. Much of the discussion reported to us was sophisticated and detailed – for example:

> *We could see that once we got children past year 3 in writing we were approaching minus 1 [on the standardised residuals] whereas with numeracy we were approaching plus 1. We said immediately:' well, if we can do that in numeracy, we should be able to do it in writing too'. Staff were amazed by that picture and writing's become a priority.*
>
> *Our entry point at the bottom end was higher than other schools we saw which means that our at-risk kids are at a high level suggesting that we have good programs in place for those kids. But we saw that the 50% group inside the block is compressed and they are exiting at a level that is not as high. It is still high, no cause for alarm, but the implication is that we should be extending our more able kids more. We've looked at that as a whole staff and everyone came to the same conclusion – we need to do something about that.*

Using the data analysis at the whole school level for school planning and for annual reports is to be expected because that was in fact the focus of the WALNA Data Club. Principals were the target audience for the Data Club. Their descriptions of the uses made of the data give a sense of the *processes* they used to engage staff and others in the school community:

> *I took it back and with the curriculum leader and the Deputy analysed for areas weakness to be addressed – in our case maths. It helped us set our priorities for next year. I also took it to staff – presented it on overheads and translated graphs and what they mean to year level teacher. I took it to the School Council – showed that we were in line and working to our target and then we worked on priorities for next year.*
>
> *We actually did our own graphing of information when it came through and we examined ourselves against the different schools within the socio-economic group and I actually spoke to other Principals in my cluster. Talked to them about what their scores were and we did a little bit of sharing around to see how we compared. And we also looked at schools within the District and looked at how we were going against that. Then we certainly used it for school reporting and school planning.*
>
> *I go over it with the School Council – they really like it as it gives them comparisons – and they like the apparent rigour. They get blown away with the graphs and boxes-and-whiskers.*
>
> *The net outcome of us talking about the data is that we are trying to be more diagnostic. We have sharpened our tools and our thinking about assessment. It's changed the conversation in the school. It's gotten us to really talk about: what is literacy?*
>
> *Besides our whole school planning, it was used as evidence for us putting in a QTP submission – we combined with another school to do a joint submission. In it we said that as a result of what WALNA data indicated we need to focus on maths.*
>
> *We used a lot of that data analysis for putting in for a literacy award this year. Found that we were able to look at K-3 and apply what we knew and our own evidence as to whether we've made a difference especially over the three years – using Year 3 data and Year 5 data. For the literacy award we were able to track and use the data extensively to show that we could really record improvements.*
>
> *At the staff development day last Friday we spent quite a lot of time thinking about the low spelling result. It stimulated a lot of really useful conversation about why we aren't achieving what we've tried to – because that's already been a priority. We decided, after a lot of conversation, to focus on only two programs that might improve spelling rather than the many different things we had been doing.*

Many principals also indicated that while they were using the data now, they expected it would prove even more useful in the future – that this is the beginning of a long-term project in the school:

> *But more importantly what we actually want to do – is have the longitudinal study – take the Year 3 and plot them from year to year from 3 to 5 to 7.*
>
> *Our capacity to make a difference will be greater in coming years with more data being available. At the moment the data to work on is limited.*
>
> *One of the things we want to do is put it onto CD – that's our plan so we can have these kids plotted and then the CD can go with them to High School.*

Interestingly, one principal who had been studying the analysis from the Data Club in detail, and finding it very informative, had decided to *not* share the information with the staff at the school development day:

> *I've got good teachers and they've been focused on the Curriculum Framework and pedagogy. My goal for that particular meeting was to organise a new structure for next year to get more collaboration. I decided that if we started exploring the data, they might look at the surface and say: the results are fine (which at the surface they are) and therefore argue against change in structure. It was a little too much like looking backwards – at that particular moment. We will all use it eventually.*

Principals were asked whether individual teachers were using the WALNA data in working with students in their classrooms – as distinct from their thinking about whole of school priorities. The answer is: not as much as at the strategic level. One respondent pointed out that the WALNA data *at the Data Club level of analysis* is not about individual students. The classroom teacher needs to go back to the original WALNA data to fine-tune planning for individual students and inform specific curriculum changes. One principal said that he gave teachers the Data Club analyses but didn't talk about any single classroom.

A significant number of principals made the point that even talking about the Data Club and WALNA data at the whole staff level was making a major contribution to individual teachers' thinking. Test data and their individual professional judgments do not always match and many teachers are taking this external reference point on board - not as the 'be all and end all', as the principals repeated often, but as a fact which needs to be considered.

> *For this school the data has been very useful because the data that was collected previously were these subjective judgements from student reports. When we put the hard core WALNA data against that and it showed the teachers that they had inflated expectations on where their students were at.*

> *The Data Club information was useful and there was an interesting aspect for us. The Deputy and I went along – she was concerned about Year 4 maths and language not being as good as the previous year. When the data came up, it was obvious the Year 4s were 'quite solid'. We see them always in school context – useful to be able to use this year's data base to put them into broader context.*

2.2 One school's experience

The story told by one school, while not typical, illustrates how the hard data generated by the Data Club became a strong impetus for change - as well as a celebration of improvements made from one testing period to the next. What emerged through the data analysis from the first testing period for this school were serious concerns about the school's performance in literacy and numeracy; concerns that led to much soul-searching on the part of the Principal and the staff. The case is presented below in summary.

> *The first time we got our results two years ago I cam away from the meeting thinking 'What the hell have I done here in two years … I'm an utter failure … how inept I must be as a Principal'. At the end of the day … we have a weak cohort, we're a PSP*

school. Socioeconomically we're very, very disadvantaged. Still compared to our like schools we were very poor – that's what got me. I didn't realise how poor we were. Since then (still including Ed Support kids) we have made quantum leaps with last year's data. The improvement has been quite dramatic (bar one student who dropped slightly). The students are mainly in or just in where they're supposed to be now.

To track this story in a little more detail it is interesting to follow the process engaged in by the principal and the staff.

I came back to the school and said 'this is an issue and we need to deal with it'.

My teachers have been involved in looking at data in overall terms with respect to the bands etc, but also looking at specific literacy and numeracy concerns. Through the process we re-planned our literacy program. Numeracy program also took a bit of a focus as well. Two years later the results were checked again and between the '98 and 2000 results some students had made dramatic improvements – not just a little gradual increase. It was a real celebration for my teachers.

What is particularly interesting in this case is the level and range of discussion and change generated by the test data. This involved staff working through what they were doing at the classroom level and the school re-prioritising and restructuring at the whole school level. It also led to broader consideration of testing and assessment and the role of schools in providing more general welfare support to students.

We went through the whole gamut of talking about teaching to tests etc, believe me. I sat with my teachers and said that in a short timeframe we can make those results go up just teaching to those things that are going to come out. But that's not the game! The game is that we have to make some real broad gains. In a very short time we went through all those things that you read about in the papers. And we had a lot of discussion about what are the real shifts we want to make for these kids.

The impact it had on our school has been dramatic - from the day I got those results.

Not surprisingly, the staff – working in a difficult school context – looked externally for explanations of why the school was under-achieving in literacy and numeracy, but in the end had to accept that they had to take up the challenge of making a difference for their students.

People went through the gamut of … 'these kids are weak and that's acceptable' ... 'we should be a Band 1 not a Band 2' – trying to explain it away. But despite all those things at the end of the day our kids end up not reading well – everyone knows that. 'So what the bloody hell are we going to do about it? You can explain it till the cows come home, but they need to have a certain level of skill.'

We came up with early intervention screening from 5 year-old program – lots of interventions. We adopted very explicit approaches to language. We changed support structure for the whole school. There was shuffling of resources. I guess it's a package over time … there are breakfast programs because a large core of kids don't have breakfast or lunch. We also provide 'feel good about myself' type of things too for kids

who come from very dysfunctional homes. We can't say it's not our role – they still turn up at school hungry and not being to concentrate. It's not just the WALNA stuff – we've been collecting data on how much food these kids are eating etc – through other surveys.

2.3 Use of the data Club at District level and elsewhere

We spoke to two District Directors and officers from two other Districts. The Directors confirmed that principals use the Data Club analysis of the WALNA test results to discuss with their District Directors what they were doing in their schools and 'how they are going'. In turn, the District Directors do specifically ask about WALNA data:

I use it in my discussions with the principals. I can talk to them about how they compare – indeed I'm doing that twice a week now as the reviews are in progress. They're comfortable showing it to me - there hasn't been anyone reluctant to show it to me. They usually say it has confirmed their judgement and, then, that this is what they are planning to do.

The conversations with principals have definitely improved since the Data Club. They're getting better at talking about the data and, especially, about interpreting it. I am sure that those in the Data Club are better at making judgments and planning. Without the Data Club, they think they're all doing okay.

It was clear to us that, in talking to a few people in District Offices, their own attitude to the data has helped to make the conversations with principals productive. It is clear that they are not using it in a harshly comparative or punitive way. They report that they do push principals to genuinely confront the data: saying, on occasion: "I don't think you're doing as well as you think…", "I think you've misinterpreted that…", "what are you going to do about…?". But they follow up with "what kind of support do you need to do…?". As one explained: it is not the game of hide-and-seek it once was with superintendents, but show-and-tell.

The sense from District officers, perhaps even more than from principals, is that the kind of analysis and professional development provided by the Data Club will be used more and more in the coming years. One particularly emphasised the importance of the value-added data, and that has hardly been used at all thus far.

One said that the analysis of the data undertaken by the Data Club has been an immense help to them at the District level:

We wouldn't have done it – mainly because I don't have sufficient people to run it – to collate it. It's a very good resource. It's just a huge deal to collect, collate – we don't have the resources available to actually collect each schools data; don't have the data base; don't have lots of things. And we also don't have the expertise. If it hadn't been put forward we probably still would not have done much with the available information. Also we felt very constrained by FOI.

One of the important problems the Data Club solved was the fear – especially amongst District and Central Office managers – that Freedom of Information (FOI) would give the media access to the WALNA data and lead to league tables.

While the Department was involved in the direct negotiations with the media, system managers praised the Data Club for further quarantining the results. One principal, in fact, recounted going to the first Data Club professional development session armed with questions about FOI and coming away convinced that the Data Club had got the problem "sorted".

The District Directors had a special session in which they went through the same sort of professional development (PD) the principals were given. Other District Office staff have attended PD sessions with principals. As a consequence, the Districts understand quite well how the schools can (and should) use the data. But at a higher level of aggregation, WALNA data provides information about the performance of whole districts. It was brought to our attention that it might be useful for the District Directors to receive PD from the Data Club specifically directed at their own use of the data.

We interviewed individuals from two school systems in the Eastern States. Both had contacted Bill Louden and Helen Wildy (creators and managers of the Data Club) because they were interested in doing something similar. Both were immensely impressed with the response and help they received. Both are determined to follow through.

2.4 Use of the website

The use of the website is varied. Some schools used it often and with confidence, others had it marked as 'favourite', yet others had not accessed it all. Several who had wanted to use it had had difficulties.

> *I certainly do use it. The only thing I'd like to be able to do is tidy up the graphs to get rid of some of the stuff around them – print bits – and be able to transport it to Word or something. Graphs without print buttons etc.*
>
> *Only difficulty is trying to access the site where we can put in data and get our own box and whiskers graph. First time wasn't on yet. Second time I couldn't get in. I rang and spoke to someone. There is still a problem with user-friendliness of website to access data.*
>
> *Have gone into the website – but didn't get the information I was looking for – whether that was because I went in the wrong direction. Was trying to compare my data but couldn't.*
>
> *No I haven't used the website., Then again, the Internet is not hooked up to my office, which makes it that much less accessible.*

Further mention is made of the website under 'professional development'.

2.5 Summary

The information gathered from this survey about the *use* being made of the Data Club by principals (and incidentally by Districts) suggests that the data, the analysis, and the skill and understanding developed through participation in the Data Club data is being applied extensively.

It is generating school-level discussions on a wide scale. In some schools there is already evidence of significant school change being undertaken to address weaknesses highlighted by the data. What has been of most immediate value is the potential provided by the data is for schools to see their achievements in context - other like schools in the same band or more broadly schools across the School District and the State.

In the process of working through the Data Club, principals (and schools) appear to have stopped being defensive if they are not performing as well as they expected. Perhaps it is that the graphs are seductively (and indisputably) clear – there seems less argument with the data, less wish to hide from these facts. With this, there is a danger that schools may become too accepting and uncritical of the numbers. The principals *all* pointed out to us that the WALNA data is one small bit of information to add to all the other sources of insight about their students and their school. Nonetheless, most schools are clearly taking the Data Club analyses very seriously.

Beyond what emerged as an extremely positive view of the usefulness of Data Club data analysis, a few concerns were mentioned about the relevance of the data for very small schools or schools with transient populations, time lags in delivery of data, and further scope for professional development. these are discussed in later sections..

3.0 PROFESSIONAL DEVELOPMENT PROVIDED BY THE DATA CLUB

The professional development (PD) sessions were extremely favourably 'reviewed':

> *Very well done – you'd have to be thick not to grasp it from the way it was presented.*
>
> *Very professional – the messages came across very clearly.*
>
> *Just what we were looking for – found it opened my eyes statistically to what we were doing and it was very easy to understand.*
>
> *The presenters had genuine authority – and they were very concerned that the message was getting across.*
>
> *It was good. really good – excellent to understand how the data was produced, what the whiskers mean. And I did understand it.*
>
> *We have 58 kids in the school… so we really had to understand it and not get mislead. Small schools must use this data differently. That was all made very clear.*

What was particularly good at the PD was the list of things to look for in the data – e.g., if you look at that percentile, you can gauge x, y…

Taking us through a graphing exercise where we did the box-and-whisker by hand helped us understand what it meant – so it was a good level of concept development.

I enjoyed it.

And so on. The one further feature of the PD noted by many of the principals – and it would distort the findings if we did not point this out – is the exceptional skill Helen Wildy brought to the sessions. Bill Louden also received praise, but the appreciation of Helen's talent as a teacher and the clarity of her explanation was considered outstanding.

Most principals attended the PD themselves accompanied (usually) by a deputy principal – although occasionally an interested teacher went instead of the deputy and, in one case, instead of the principal. A number commented that it was a very good idea to have two people from each school – by talking to each other about their data during the PD session, they explored and interrogated it more fully.

While the principals were well satisfied with the PD they received many of them would like the Data Club to have an on-going cycle of PD. There are three main reasons for this.

- more PD as a refresher for the principals. One principal was thinking out loud about this in the interview:

  I'll send someone else next time… no, I won't. I would still like to go, to reconfirm what I know and there are always new bits to understand – and it is so easy to misrepresent it.

  Have done it twice now. I found the second time around was much more meaningful – good for me to be able to have a refresher.

- new PD because as the 2001 (and subsequent) cohort data comes in, schools will be needing to deeply understand how to analyse and think about tracking student performance over time. They are not confident they have received enough specific skilling in that area (as the data has not really been available) - *I'll try to use the 2001 data on my own, but I'm not confident.*
- more PD so that more people from each school could develop competence. One respondent pointed out that principals change schools with some degree of frequency – more than teachers in most cases. Thus, it would be wise for knowledge about using the data (and an organisational memory of the data) to be embedded in each school.

There was some comment on the number of participants in each PD session. The number appeared to range widely. We did not specifically ask about the size of the groups but one person mentioned being in a group of 10 for one PD session and in a group of 30 for another. Thirty seemed too high to that principal.

A few other specific suggestions about enhancing the PD were made:

- maybe they could do the PD at SIDE where there is a computer lab and actually get people there punching in information and seeing what comes out;

- more training on using the website tools – especially for getting graphs to use in presentations;
- more help in making a direct link between the raw score and the outcome data. As it stands now there is just a pictogram and need to approximate it across. Wants a table like the one that converts raw scores to WAMSE;
- possibility of special PD for small schools – although one of the principals from a very small school (20 students) said he preferred doing the PD with the whole range if principals;
- possibility of special PD – or more specific help – for schools with large transient populations. Principals are concerned that they may be comparing different students in, for example, 2001 Year 5 data than the children who were actually in Year 3 in 1999.

In sum, the professional development aspect of the Data Club has provided needed tutoring about analysing and using the WALNA data in schools, and it has been very much appreciated. There was not a single principal who felt that he or she had not learned what was intended for them to learn. The outcome from such successful PD is that the principals want more – more for themselves and for their teachers.

A number of principals were rather unsure about what they could or should expect *next* from the Data Club. This may be information which could be posted on the web site.

4.0 RELATED ISSUES

A number of points were raised in the interviews which are relevant to the Data Club:

4.1 Time between testing and their processing

The longer it takes for the test results to be put into the hands of principals and teachers, the less its value. This is especially true for teachers who would still be teaching the children who had been tested if the tests were processed quickly.

> *The results come out at the end of the year and they don't get picked up really till beginning of the next year. By then that group of kids has moved on and teachers think it doesn't really effect them..*

Many principals asked us, plaintively, whether we knew when the 2001 results would be available. They want it. Whether it would solve the time delay problem if the testing people understood that the data really are being used, it may still be a good idea to remind them that people are relying on it. Real deadlines should be imposed on the process.

4.2 Confidence in the testing regime

We specifically asked principals whether their participation in the Data Club had increased their confidence, and the confidence of their teachers, in the WALNA testing regime. A few schools remain firmly opposed to the tests but there does seem to have been a significant shift in opinion:

Like most people I was a bit 'suss' about it originally but I think no-one can question the data that's coming out. And I think that's spread through the general teaching staff. Certainly the process we used – looking at the WALNA data and teacher judgements - really made them re-think and re-focus.

The Data Club has given me a better explanation of what's been gathered. It's not the be all and end all, but it has a contribution to make.

I think it has helped convince teachers they need that kind of information as well as teacher judgment. Teacher judgment used to be all we had.

The first year all of our teachers refused to do the test but now a couple of teachers are doing it saying, 'what the heck, we want to learn from it anyway.'

Attitudes haven't changed yet in this school, but I anticipate that this will happen within 18 months.

I think confidence is improving. Part of it is the kids are getting used to testing. It was not something they had done before. Going through the test situation is important for kids so that they are not thrown by it. Wasn't part of their culture before.

Some concern was expressed that schools are teaching to the test – as the above quote indicates, and at least one other principal mentioned that in his school teachers were giving kids a number of different reading, writing, spelling and maths tests to familiarise them with the kinds of questions they might be asked. A third principal made the point that even if they were 'teaching to the test' in some way that may be no bad thing

With outcomes and no syllabus, it is easy for teachers to miss out chunks. Using WALNA as a cross-reference helps. In fact, this year a teacher who had been opposed (and not run it), did and said, 'it's helped me to see what content they are tested on'.

There were patchy concerns about the test conditions not being consistent across schools - especially where the normal classroom teachers refuse to run the tests and outsiders, who may know nothing about the students and little about the tests are hired. There were also questions asked of us as to whether schools were counting or excluding the results of 'Ed Support' students and whether the 'H factor' truly generates a band of like schools.

These concerns are relevant to the Data Club in that it may be part of its professional development role to specifically think about shoring up teachers' confidence in the tests. It was certainly the case that principals take Louden's and Wildy's involvement in the Data Club as a kind of imprimatur for the WALNA tests.

4.3 Data Club in the hands of independent and highly respected educators

The standing of Louden and Wildy, especially amongst school principals, has contributed in a major way to its take-up. From the responses of principals, there seems no doubt that something like the Data Club needs to be outside the school system.

We did not ask principals if they knew the funds which enabled the Data Club to be formed and operate came from DETYA. We suspect that having the funding come from outside the school system also counters schools' natural suspicion of things that come from central offices.

With the exception of the promptness with which the schools obtain their data, the other issues raised are, at their core, about the trust and confidence school principals and teachers (and ultimately parents) have in (1) the testing regime, (2) in one another, and (3) in the people managing the Data Club. Continuing to improve that trust and confidence in the case of the first two and maintaining the exceptional high level in the third is important if the WALNA data are to lead to improved outcomes for students.

The Data Club has begun very well, but its role has only just begun. Schools recognise that there will be much more for them to learn about using the data over the next few years. And they will want reliable help from independent experts. The Data Club has provided those services to everyone's satisfaction – indeed, it seems to have exceeded expectations.

Appendix 2: Literature Review and Data Display Options

Developing Schools' Capacity to Make Performance Judgements:
Literature Review and Data Display Options

September 2001

William Louden and Helen Wildy
Institute for the Service Professions
Edith Cowan University
Perth, Australia

Contact:
w.louden@ecu.edu.au
http://isp.ecu.edu.au/dataclub

TABLE OF CONTENTS

TABLE OF FIGURES

TABLE OF RECOMMENDATIONS

1. In the absence of adequate time-series data from WALNA, schools make use of like-schools analyses.

2. As adequate time-series data become available from WALNA, schools make use of value-added analyses.

3. The Ross Farish H index be the area-based measure of SES used in WA like-schools calculations.

4. The adequacy of 'H' in comparison with individual measures of SES such as parental education and income be investigated in a sample of WA schools.

5. Consideration be given to the introduction of unique student identifying numbers, to improve the precision of like-school and value-added calculations.

6. In the absence of unique student identifying numbers, simple regression rather than multilevel modelling be used to make like-school and value-added calculations.

7. Statistical uncertainty as well as calculated school values be represented in like-school and value-added data displays.

BACKGROUND

The introduction of national population assessment in Years 3 and 5 has changed the educational landscape in states such as Western Australia. Until 1998, literacy and numeracy assessment conducted under the Monitoring Standards in Education was based on Statewide sampling of students' performance. Through this program it was possible to generate Statewide estimates of the range of performance at each grade level and of the average performance in the grades and learning areas assessed (see, for example, Ministry of Education, Western Australia, 1991; Education Department of Western Australia, 1995, 1999). In addition, it was possible to provide estimates of the relative success of different social and cultural groups. What was not possible, however, was to provide information to schools about either the progress of their whole cohort (since not all students were assessed) or their relative progress of the whole cohort compared with students in other schools. So, although the sample assessments were useful for indicating broad Statewide trends, they were not a useful tool in school improvement.

The combination of improved whole population data through the Western Australian Literacy and Numeracy Assessment (WALNA) program, and an environment of increased school-system interest in accountability has led school principals to be more concerned to learn how to use their students' performance data. Technically, the obstacle to using student performance data for school improvement is that it is difficult to be sure how much of any reported student performance can be attributed to the efforts of schools. There are two sources of uncertainty: the home background of students, and the prior achievement of students. Because the correlation between socio-economic status (SES) and school mean scores is typically about 0.7, about half of all the variation in school mean student scores can be predicted by home background (Marks et al, 2000, p. 36). Similarly, about half of the variation in students' scores can be predicted by prior achievement (Fitz-Gibbon, 1997, p, 50). Together home backgrounds and prior school learning performance combine to predict as much as three-quarters of the variation in the outcomes of schooling for older students. Thomas & Mortimore (1996), for example, have reported that some 70-75 percent of variation in 16-year-olds' school performance can be explained by student intake measures in individual schools. Much smaller estimates of the impact of home background have been reported in studies using individual level data (Marks et al, 2000) or multi-level statistical analysis (Rowe, 2001).

From the point of view of school improvement, however, the key issue is to separate out intake measures such as SES and prior performance from school effects. Researchers interested in helping schools to estimate the proportion of variation that can be attributed to school controlled variables – such as teaching, the allocation of resources and quality of school leadership – have developed several analytical strategies. One strategy, usually called the *like-schools* approach, attempts to estimate relative school effects by making comparisons between schools with socially similar student cohorts. The second strategy, a family of approaches called *value-added* assessment, attempts to estimate relative school effects by measuring progress over a particular period of time. Strategies used in each of these approaches are described in the next two sections.

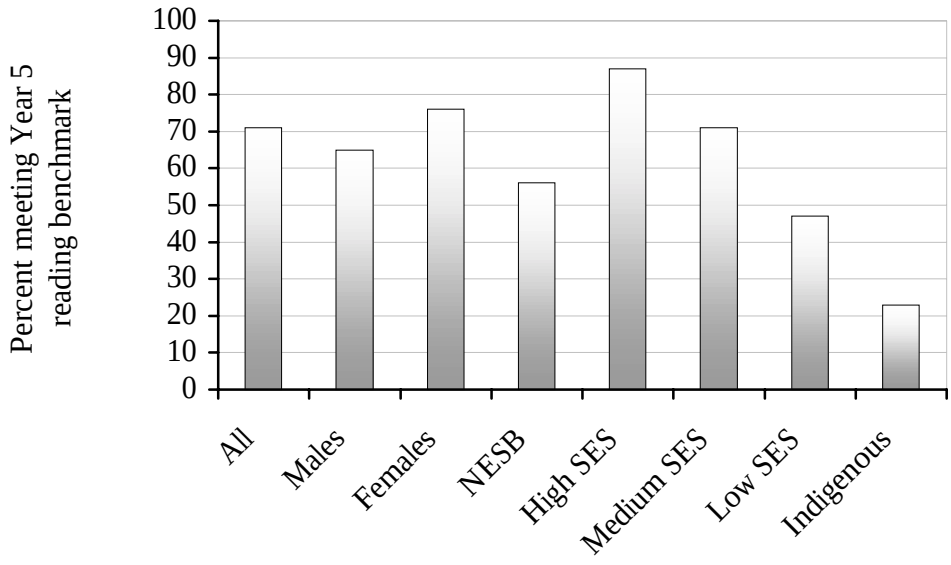
Whichever method is used, caution must be exercised in interpreting results of test programs as evidence of school effectiveness. No matter how well a particular whole-cohort testing program may measure the curriculum, a test is a sampling device. Test items sample a range of possible performances, based on considerations about what matters most in the curriculum and what discriminates among the performances of participants. Not all of the outcomes sought in a learning area are assessed, not all of the learning areas are assessed, and not all of the valued outcomes of schooling concern the cognitive domains most frequently assessed. Some tests are psychometrically better than others are, and some have higher content validity, but all testing programs contain high levels of uncertainty. In addition to uncertainty related to the behaviours sampled by the test and the variation in test conditions, there is statistical uncertainty in the results.

The next two sections of this review describe the like-schools and value-added strategies. The third section of the review returns to the issue of uncertainty, canvassing the limits to interpretation of like-schools and value-added results and provides a series of recommendations for reporting to schools. The final section outlines the data display strategies adopted for 2001 by the Data Club.

LIKE-SCHOOLS ANALYSIS

Like-schools analyses use cohort similarity as the predictor in estimates of school effectiveness. An indication of the range of difference in Australian schools is provided by national benchmarking studies in literacy. For example, the percentage of students in the National School English Literacy Survey reaching the (then) draft national literacy benchmarks varied according to gender, home language, socio-economic status (SES) and ethnicity (Masters & Forster, 1997, p. 15). In particular, the differences according to socio-economic status were stark. Whereas 87 percent of high SES students met the Year 5 benchmark, 71 percent of medium SES students and only 47 percent of low SES students met the benchmark. The results for the special indigenous sample were even starker, with only 23 percent achieving the benchmark score. (See Figure 1, below.)

Figure 1. Benchmark performance by social group, NSELS, 1996



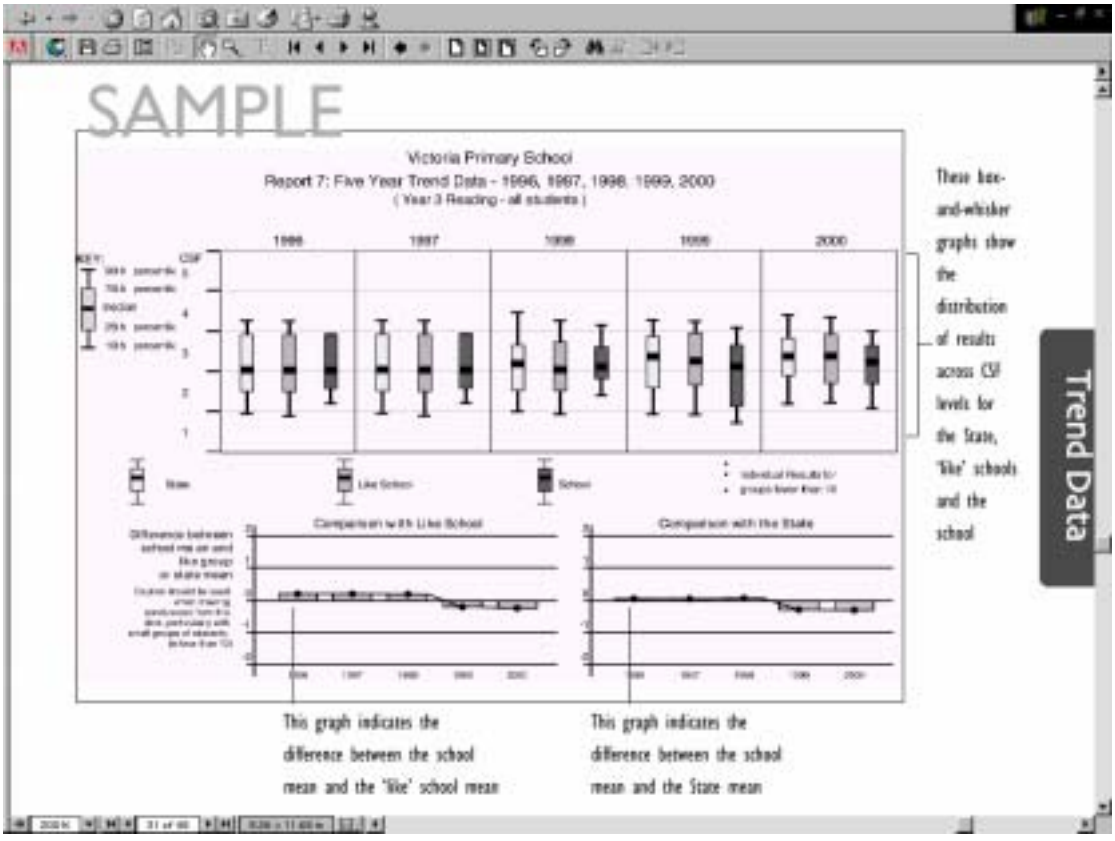
Provided that schools each shared a heterogeneous group of children, mixed equally in terms of race, gender, home language and SES, a like-schools analysis would not be necessary. Australian school populations are, however, sharply divided in terms of the social circumstances of the communities they serve. Through schools, these differences leads to what Teese has called 'the social geography of success and failure' (2000, p.35).

One of the Australian schools authorities that has attempted to account for the social geography of schools through a like-schools approach is the Victorian Board of Studies (2000). Through the Achievement Improvement Monitor (AIM) program (http://www.vcaa.vic.edu.au/aim/), results in the full cohort Statewide testing at Years 3 and 5 are reported for individual students, for each school, for like-schools and for the whole state. Among the analyses returned to schools is the data display shown in Figure 2, below. For this sample school, the display provides a five-year trend in scores in Year 3 reading. For each year, a box-and-whisker plot represents the range of performance against a scale provided by the State's Curriculum and

Standards Framework (CSF) levels. As is the statistical convention, the central shaded 'box' in each plot represents the middle 50 percent, between the 25th and 75th percentiles, and the line towards the middle of the box represents the median score. The 'whiskers', the horizontal lines above and below the central shaded boxes, represent the 90th percentile and the 10th percentile of student scores. In this sample school, in 2000, the school's median performance was a little below the State and the like-school scores, as was the performance of the 10th percentile, 25th percentile, 75th percentile and 90th percentile. Compared with the performances recorded for 1998, where the 10th percentile, 25th percentile and median scores were all above the like-school scores, this would represent a disappointing level of relative performance for the less-accomplished half of the school cohort.

The line and column graphs below the box and whisker plots in Figure 2 provide a time series analysis of school, like-school and State means – expressed in standardised residuals. Compared with like-schools, the sample school has had residuals of the order of +0.2 for the years 1996-1998, and residuals of the order of -0.2 for the years 1999-2000. Compared with the State, residuals of 0 in 1996-1998 have been followed by residuals of about –0.3 in 1999-2000. Together with the evidence from the box and whisker plots, the total fall of about 0.4 standardised residuals compared with like-schools and 0.3 compared with the State might indicate some cause for concern in the sample school.

Figure 2. Data Display: AIM program, Victoria

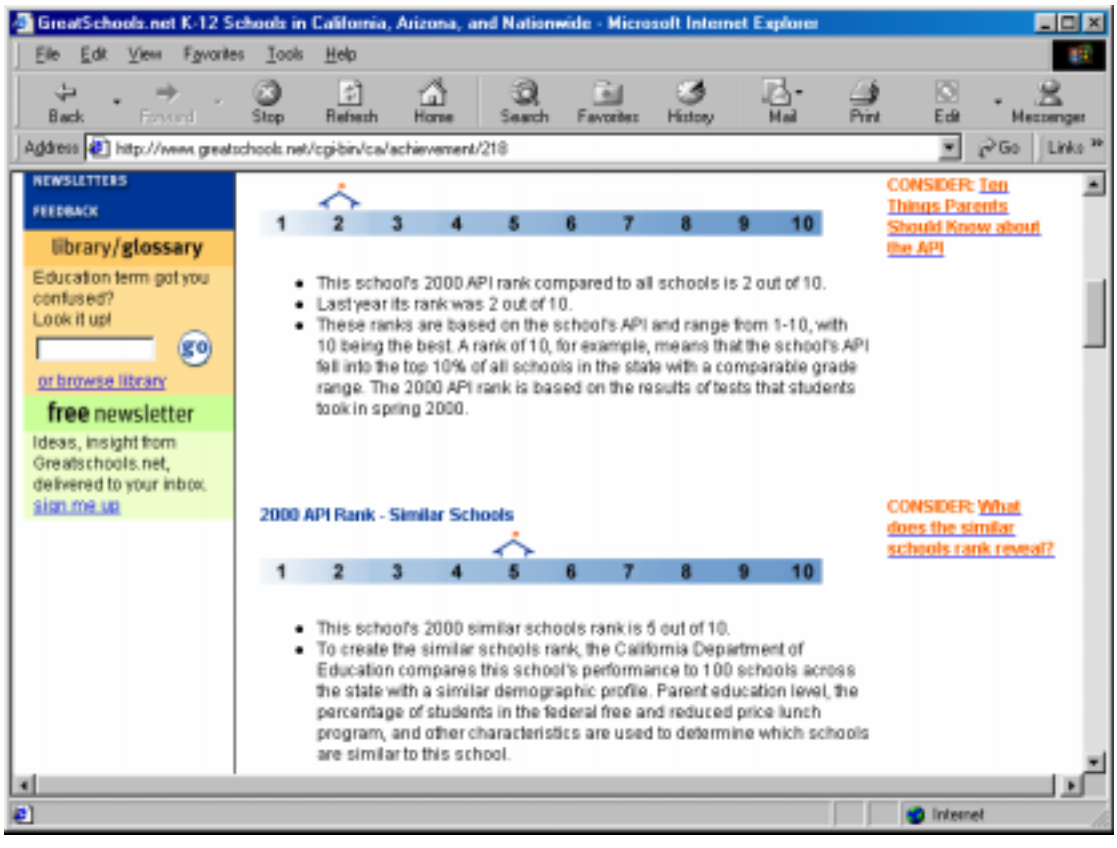


The like-school approach to accounting for differences in school intake variables is perhaps most comprehensively developed in California, where a program of annual, Statewide standardised testing is linked to an elaborate index of school similarity (http://api.cde.ca.gov/). There are two parts to this like-schools procedure, the Academic Performance Index (API) and the School Characteristics Index (SCI)

The API is a composite measure of student performance, based on student SAT 9 scores in the Standardised Testing and Reporting (STAR) program (http://star.cde.ca.gov/). The California Department of Education provides an API ranking for all schools, dividing all schools within a phase of schooling into decile groups. School performance is then reported in terms of API ranking.

In addition, California has adopted a complex School Characteristics Index (SCI) which provides a composite index based on background characteristics including student mobility, ethnicity, SES, home language, class size and teacher credentials. Derived though multiple linear regression, the SCI groups schools with similar levels of advantage or disadvantage in terms of factors affecting student performance. Within a phase of schooling (elementary, middle or high school), schools are grouped with the fifty schools above and below their SCI score, providing a comparison group of the 100 schools most similar, sorted by phase of schooling. The 100 schools are then separated into deciles according to API. The combination of SCI and API allows the Department of Education to report an API ranking for Similar Schools. Within a school's 100-school group, schools are ranked in decile groups from 1 to 10, with deciles 1 or 2 rated as 'well below average for similar schools' and deciles 9 and 10 rated as 'well above average for similar schools'. For each school, the 100 closest SCI scores comprises a relatively narrow range, about 2 percent of the State's elementary schools (Rogosa, 2000, p.10)

Figure 3. Third party reporting of California API and API Similar Schools ranking



The combination of API ranking and API Similar Schools ranking provides two estimates of school performance, one in comparison with all schools and one in comparison with 100 similar schools. This allows a graphic display such as Figure 3 to be provided to the community. In this case, the information is provided through a third party web-site (www.greatschools.net), and refers to a school serving a poor community in Oakland CA. Although the school's API ranking is a low 2, the API Similar Schools ranking of 5 rates it as about average in comparison with schools serving similar communities.

> The strength of like-schools analysis is that it appeals directly to teachers' sense that it is easier to produce high test scores in schools located in what are euphemistically called the 'leafy suburbs'. If there are to be school comparisons based on student performance, it may be argued, at least the comparisons should take into account some of the student intake characteristics that predict higher or lower levels of performance. The next section of this paper describes value-added analysis, an alternative strategy that attempts to account for likely differences in terms of prior performance, rather than the social characteristics of the school intake.

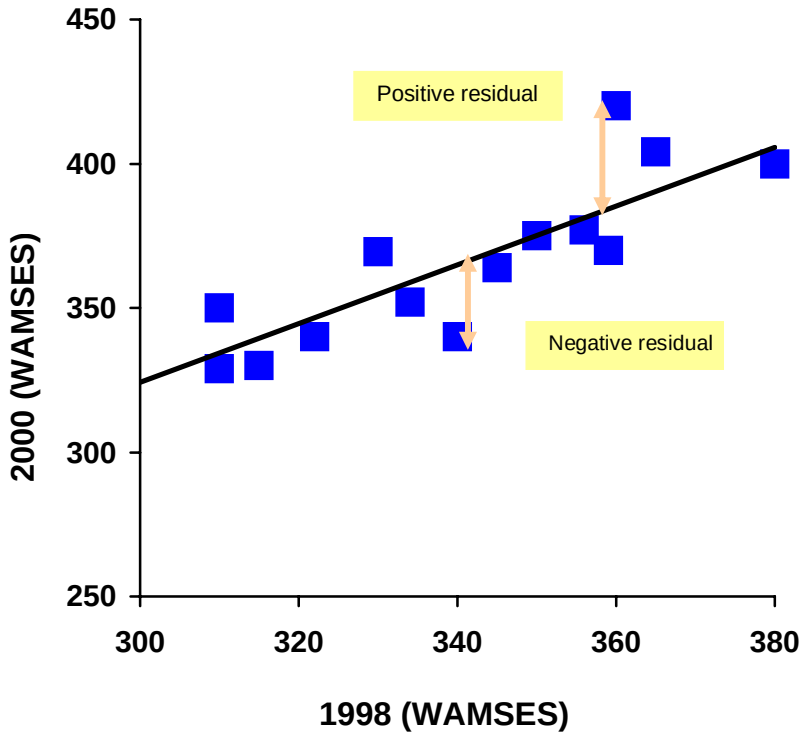
VALUE-ADDED ANALYSIS

Value-added analyses use prior attainment as the predictor in estimates of school effectiveness. Instead of the performance question being phrased as 'How well did students in this school go compared with other schools', the question is, 'How much did students scores improve over time compared with the expected rate of improvement?'

Carol Taylor Fitz-Gibbon, whose A-level Information System (ALIS) value-added program the pioneered the British use of value-added measures of school effectiveness, describes the analytical process as having four steps (1996, pp. 120-128). Data are collected (1) on prior achievement and (2) subsequent achievement by the same students, (3) the two data sets are related, and (4) the relationship between the two is used to predict 'the achievement that might reasonably have been expected' (p. 121). The strategy for relating the two data sets is simple regression analysis. The difference between the achieved scores and the expected scores (represented by the line of best fit) is called the residual. The residual, which may be positive or negative, provides an estimate of value added.

Figure 4 provides a local example of the value-added strategy using 1998 and 2000 mean scores in reading in for a sample of schools, on a common scale, the line of best fit, and the positive or negative residuals.

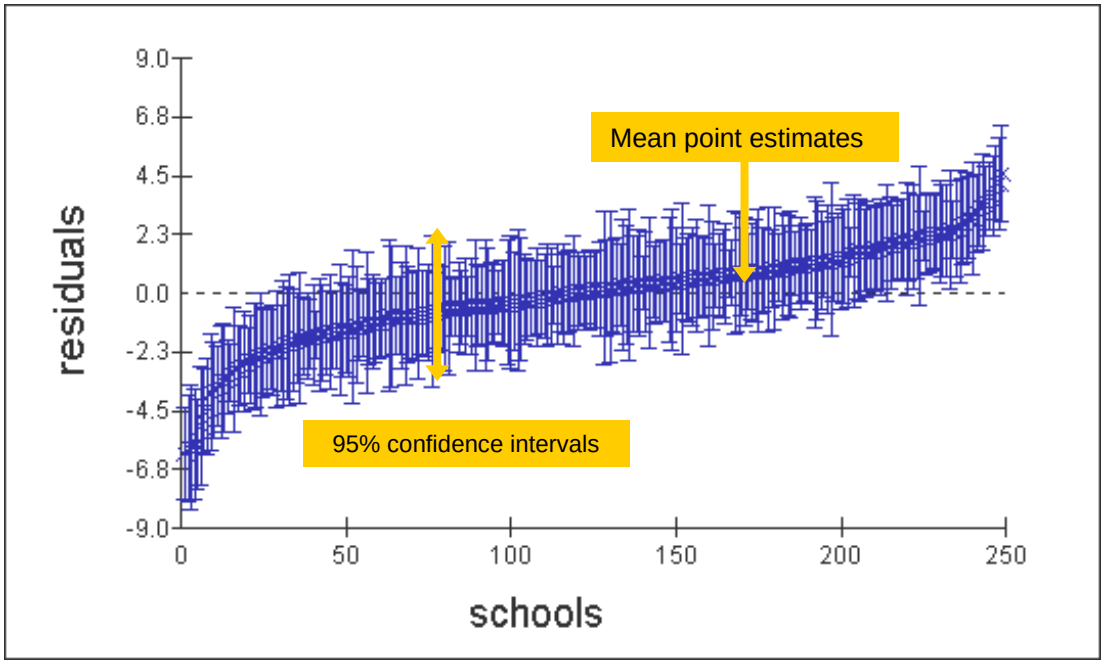
Figure 4. Value added analysis



Each of Fitz-Gibbon's four steps to producing value-added estimates can be traced though Figure 4. For this group of 16 schools, the horizontal axis describes prior achievement (Step 1); in this case 1998 mean school reading scores for Year 3. The vertical axis describes subsequent achievement (Step 2), the 2000 mean school reading scores for the same group of schools. Performance is measured in both cases on a common scale (Goldstein, 2001), called the WAMSE scale, which allows comparisons to be made between WALNA reading test scores over time. Fitz-Gibbon's Step 3 is to estimate the relationship between the sets of prior and subsequent scores, represented by the black line of best fit. This line, the regression line, predicts the expected score in 2000 for each achieved score in 1998. The residual – the difference between any one school's expected score and the achieved score in 2000 – is the estimate of value added (Step 4).

An alternative strategy is to represent value-added scores for a group of schools such as those in terms of standardised residuals rather than school mean scores. Figure 5 (based on Rowe, 2000, p.81) plots the residuals for 250 schools in rank order. For each school, the mean point estimate of the residual and the 95 percent confidence interval associated with that estimate is represented. If any portion of the confidence interval overlaps zero, then there is no statistically significant difference ($p>0.05$) between the school's expected and achieved scores; that is, student scores have improved as much as expected. In schools whose lower confidence interval is greater than zero, students have improved more than expected and the school may be said to have added more value than other schools. In schools whose higher confidence interval is less than zero students have improved less than expected and the school may be said to have added less value than other schools in the analysis.

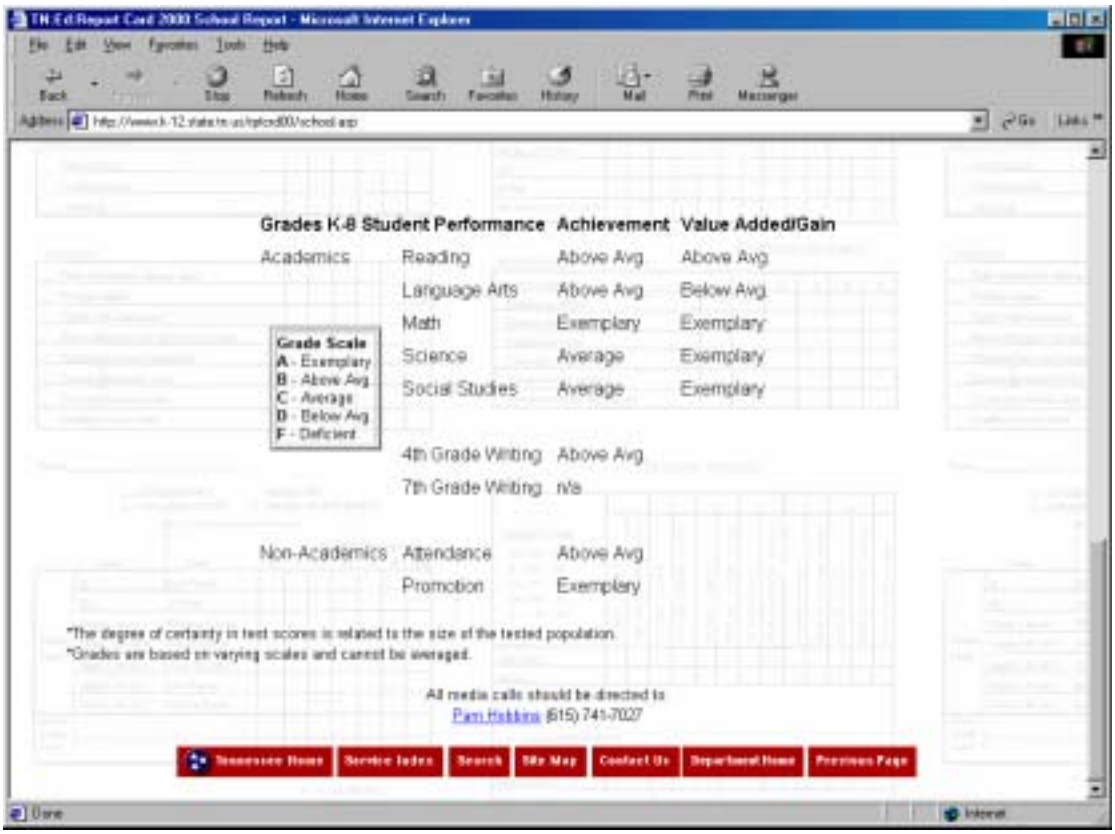
Figure 5. Residuals: Mean point estimates and confidence intervals



Whereas the English value-added system has emerged from more than a decade of bottom-up cooperation between researchers and schools, the most prominent value-added systems in the United States have been top-down systems mandated by law. For example, value-added calculations on all schools routinely are provided by law in the State of Tennessee, through the annual Report Card based on the Tennessee value added assessment system (Sanders & Horn, 1994; Sanders, 1998). State law requires the State Department of Education to report on the extent to schools and districts meet a set of minimum expectations, and a set of goals to be attained. Among the minimum expectations, schools and districts are required to demonstrate that students gain a year's average growth compared with national norms. This gain is expressed as a 100 percent value-added gain. Schools and districts are rated 'A' and 'Exemplary' if they achieve scores of 115 percent, or 15 percent more than national norms for growth. Similarly scores between 105.0 and 114.9 are rated as 'B' ('Above Average'), scores between 95.0-104.9 are rated as average, and so on. Scores below 84.0 are rated 'F' and 'deficient'. Public access is provided to school in the form of raw scores and report card style value-added results.

Figure 6 shows the value-added report card for one such school, Briceville Elementary. In this school, although language arts scores were above average, the value added to language arts scores sine the previous year was a D, average, while science and social studies scores had shown more growth than expected and are reported as Exemplary. (see: http://www.k-12.state.tn.us/rptcrd00/default.asp)

Figure 6. Tennessee value-added report card, Briceville Elementary



In addition to this simplified report card form, the Tennessee system provides detailed year, by year breakdowns of school and district performance in each of grades 4 to 8, in each of the learning areas assessed (see: http://www.k-12.state.tn.us/rptcrd00/). School-by-school reports are published by law in the highest circulation newspaper in each school district and a 'user- friendly' version of the results also appear in the Nashville *Tennessean* newspaper and are available from the newspaper's web-site (http://www.tennessean.com/schools/).

The strength of value-added analyses is that that they make more appropriate comparisons than like-schools analyses. Instead of estimating the impact of background variables such as SES on subsequent achievement, value-added analyses deal directly with the relationship between prior and subsequent measures of student achievement. Both value-added and like-schools measures, however, depend on a chain of interpretation and representation which can influence the kinds of conclusions drawn from the analyses. These issues of interpretation and representation are taken up in the next section of this paper.

ISSUES IN INTERPRETATION

Like-schools and value-added comparisons share the common goal of separating out home effects from school effects in student performance. Because home background and school experience interact in complex ways in each school student's education, and because phenomena such as SES and student achievement are social constructs, the best that measurement can do is provide estimates performance. The foundations of such estimates are in student assessments, which are no more than samples of individuals' possible performance. For some students, favourable combinations of home background, school experience and curriculum contact will combine to produce very high measures of performance. For other students, much less favourable combinations will produce very low measures of performance. As such test scores are displaced from individual students to school distributions of scores, summarised in terms of school means and standard deviations, and transformed by regression against measures of social class or prior performance, the possibilities of error grow. Such error cannot be eliminated. Inevitably, it accompanies statistical manipulation. The best that can be done is to be mindful of potential sources of error in any set of interpretations. In the following section, five issues of interpretation and representation are considered.

1. Choosing between like-schools and value-added analysis

Like-schools analyses adjust for intake characteristics and value-added analyses adjust for prior performance, but which provide the most useful information for schools in managing their own planning? One response to this question, drawn from analyses by Goldstein and his colleagues, is to ask whether the choice of method makes any difference to the outcomes of analysis.

Figure 7 below, reproduces a figure from Goldstein (2001, p. 9). The three graphs represent residual plots with 90% confidence intervals for three mathematical models.

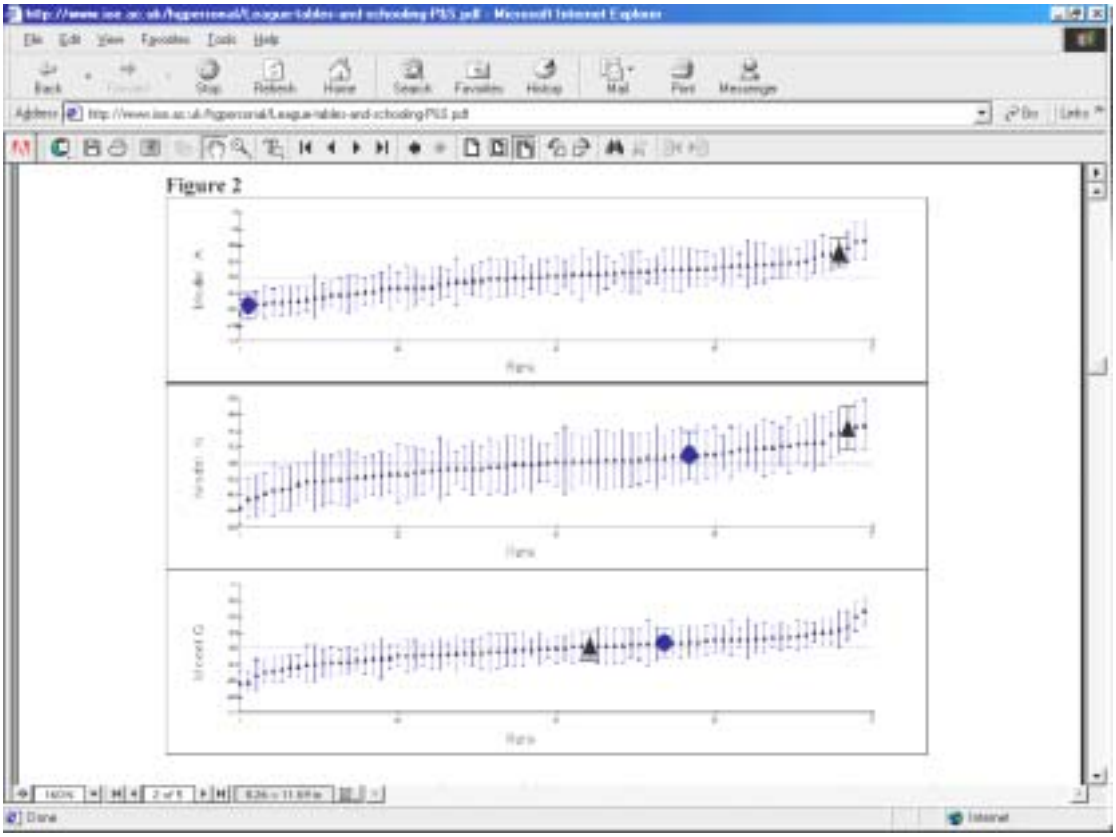
- Model A is the set of residuals from raw Key Stage 2 mathematics test results for a group of school ranked with 90% confidence intervals. This procedure suggests that based on raw scores, the confidence intervals for most schools overlap zero, that scores in the school marked with a diamond are lower than expected and that scores marked with a triangle are higher than expected.

- Model N is the set of residuals based on a regression that adjusts for free school meals at an individual and school level. Taking into account impact characteristics, this like-schools analysis suggests that the school marked with a diamond is performing as expected and that the school marked with a triangle is performing better than expected.

- Model Q adjusts for free school meals and for value added since Key Stage 1 assessment in mathematics. Neither the diamond school nor the triangle school has residuals different from expected.

That is, despite the significant difference between both schools in the raw score model and the significant difference from the expected result in one school on the like-schools model, there is no statistically significant difference between the schools in the like-school plus value-added model.

Figure 7. Raw score, like-school and value-added ranking



On the basis of Goldstein's analysis, it may be concluded that the choice between like-schools and value-added estimates is not trivial, and can lead to diametrically opposed conclusions. In choosing between these strategies, the consensus in the literature is that value-added models provide superior estimates of school effects. Sanders, who began the US value-added research during the early 1980s has reported that (for grades 3 to 8 in Tennessee) cumulative gain in student performance was 'virtually unrelated' to mean school achievement level, percentage of students receiving free and reduced prices lunches, and racial composition of the school (Sanders, 1998, p.26). On this basis, he argues, it is unnecessary to include background variables in value-added measures of school effectiveness. Moreover, as Fitz-Gibbon (1996) has argued, adjusting for SES can be seen as an 'excuse':

Teachers do not wish to be told to expect less because a student is from a low-SES family. Teachers take students as they find them and the expectations for student achievement can only be based on past achievement, not home circumstances (p. 148).

In the Western Australian context, however, the current WALNA data are better suited to like-schools than value-added analyses. The set of data available includes Year 3 reading from 1998, and Year 3 and 5 reading, writing, spelling and numeracy from 1999 and 2000. The only comparable set of data available that allow for calculations of value added are the 1998-2000 reading scores. For this reason, like-schools analyses have been conducted in Round 1 and Round 2 of the Data Club. Round 2 also included a single value-added analysis based on the 1998-2000 reading scores. In 2002, a more adequate set of time series data will allow calculation of value added in readign, writing, spelling and numeracy from 1999 ‘s Year 3 assessments to 2001’s Year 5 assessments.

Recommendations:

1. In the absence of adequate time-series data from WALNA, schools make use of like-schools analyses.
2. As adequate time-series data become available from WALNA, schools make use of value-added analyses.

2. Representation of socio-economic status

Like-school measures require the construction of an index of similarity, which usually includes some measure of individual or school average socioeconomic status. Such measures may be taken at the level of individual families, or based on the areas in which children live or go to school. Individual measures of SES – parent or guardian occupation and educational background – have been recommended by the recent report on measurement of SES prepared for the National Education Performance Monitoring Taskforce of the Ministerial Council on Education, Employment, Training and Youth Affairs (Marks, McMillan, Jones & Ainley, 2000). Although such individual level indicators have been available to researchers by negotiation with schools and parents, individual measures of SES are not held by the Education Department of Western Australia or linked to WALNA assessment data.

In the absence of individual data, there are several options for school level estimation of average SES. One measure frequently used in the UK and USA is eligibility for free or reduced cost school meals. In the absence of such entitlements in Australia, a similar measure is the proportion of families with health care cards. Although this data is collected by some State education departments as an estimate of relative disadvantage, it is not available in Western Australia. An alternative strategy is to use census-based data. The Australian Bureau of Statistics (1998) *Socio-Economic Indexes for Areas* (SEIFA) is a composite measure different of socio-economic conditions by geographic areas. SEIFA includes weightings for income, wealth, occupational status, level of education and other differences among communities. Although SEIFA is available at the Collection District (100 dwellings) level, mapping of SEIFA scores to school boundaries would be a laborious task. In so far as the schools involved in the Data Club are primary schools serving the local neighbourhood, SEIFA would provide an adequate area-based estimate of SES. SEIFA would provide a less effective

estimate for schools drawing outside local boundaries. A second weakness of SEIFA for the current purpose is its lack of familiarity to school staff.

Among Western Australian government schools the most familiar area-based index of SES is the Ross Farish 'H' Index (Ross, Farish & Plunket, 1988; Farish, 1993). This index, which was developed in the early 1980s, has a long history of use by the Australian Government to allocate funding through the Disadvantaged Schools Program. Based on an analysis of census data at the Collection District level, there are various forms of the index with a range of alternate weightings. The modified 'H' form currently used by EDWA to allocate literacy and numeracy, students at risk, behaviour management and retention funding is based on the dimensions for occupation, income, family structure, accommodation, tenancy, English language competence and Aboriginality. Occupation and income have double weighting. The dimensions are standardised (mean = 100, standard deviation = 10) but not normally distributed.

There are several limitations to the H Index. Because data become available two years after each Census, records can be as much as six years old and do not reflect the circumstances in rapidly changing neighbourhoods. Second, changing socio-economic conditions, such as sale of public housing to long-term renters or census collection during and extended rural drought, can affect H index scores without changing the social composition of a school's intake group. Third, although EDWA calculates the index using a sample of actual addresses in each school, schools which recruit across designated school boundaries are sceptical about their H index scores. These concerns about the fairness of H scores are reflected in the recent call in the Robson review of structures, resources and services provided to government schools in Western Australia for a 'close examination' of the 'accuracy of the index in measuring disadvantage' (*Investing in Government Schools*, 2001, p.89).

Recommendations:

3. The Ross Farish H index be the area-based measure of SES used in WA like-schools calculations.
4. The adequacy of 'H' in comparison with individual measures of SES such as parental education and income be investigated in a sample of WA schools.

3. Individual and group data

Although area-based estimates have been found to provide adequate approximations of socio-economic differences in the context of research on Australian schools (Ainley, Graetz, Long, & Batten, 1995, pp. 77-90) the choice between area-based and individual measures is more than a matter of convenience. Area-based measures of socioeconomic status show higher correlations with measures of school performance than individual measures of socio-economic status. As Marks et al (2000, p. 36) note, although correlations between area-based measures of socioeconomic status and school test scores can be as high as 0.7, correlations with individual level measures of socioeconomic status are more typically in the range 0.2-0.3. That is, basing a like-schools analysis on school-level – rather than individual level – measures of SES has the tendency of overstating the impact of home background on student performance. This is an especially

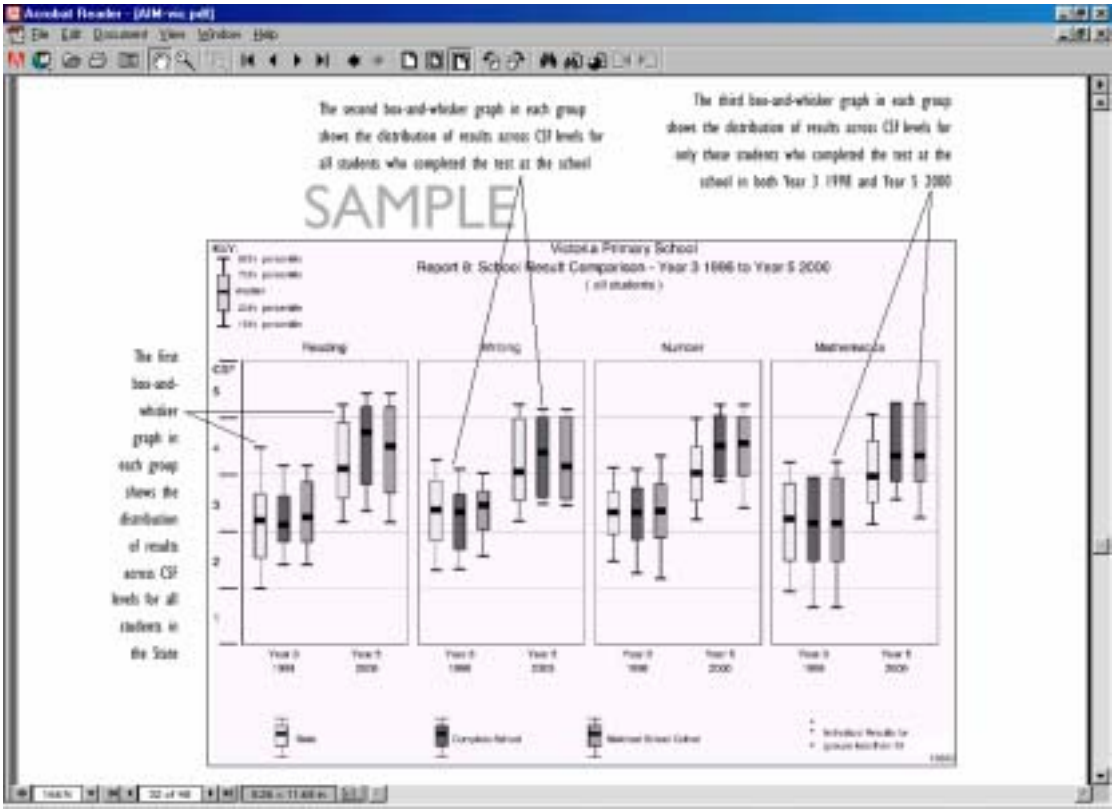
troubling possibility in schools catering for many families living in poverty, as it may mislead teachers into thinking that there is not much that schools can do about pupil performance. This is clearly not the case, as individual correlations between SES and performance in the range of 0.2 – 0.3 suggest that SES can explain only 14-17 percent of variation.

Unfortunately, as Rogosa (2000) notes, 'the considerable problems in describing individual processes (eg. student academic achievement) using group (eg. school-level data) are well-documented in every area of social science' (p. 45). Using group mean data in a like-schools analysis is likely to lead to the 'ecological fallacy' (see Fitz-Gibbon, 1996, p. 142), overstating the strength of the relationship between home background and student achievement.

In addition to the problem of the ecological fallacy in like-schools analysis based on group SES data, the quality of value-added calculations may be affected by using group measures of student performance. The WALNA data set includes individual level data in reading, writing, spelling and numeracy in Years 3 and 5 (with Year 7 assessed for the first time in 2001). Although the data are individual data, the individuals are not identified in the analysis by name or by a unique numerical identifier. Distributions of school scores are calculated, but the identities of children who contribute to each distribution are not linked to the data. For this reason, unless a school has no change in enrolments, comparisons between school distributions from one time to another are not necessarily comparisons of like with like.

In the Victorian AIM program, where individual identifiers allow for data matching, a separate analysis is available for 'complete school data', the students present in the current assessment year, and 'matched school cohort' data, the cohort of students assessed in both Year 3 and Year 5 in that school. The data display for a complete school data appears as Figure 8. Schools are thus able to consider their results including the current group of students, or in terms only of the students whom they have had in two consecutive assessment periods.

Figure 8. AIM analysis for whole school cohort



Although overseas evidence suggests that the individual effect of transience is small (Fitz-Gibbon, 1996, p. 52), both the number of schools attended and the length of time in the final school are associated with student progress (Yang, Goldstein, Rath & Hill, 1999). In Australia, very high levels of transience are associated with frequent absence from school and very poor school performance, particularly among indigenous children (Hill, Comber, Louden, Rivalland & Reid, 2001, pp. 81-84). Without individually identified data and a data matching process, transience can be accommodated in a like-schools analysis. Schools with high levels of transience are likely to be associated with low 'H' scores, and be compared only with other schools with low 'H' scores and (typically) high levels of transience. In value added calculations, however, the lower levels of attendance associated with high transience reduces children's opportunity to learn, and is likely to lead to lower levels of value-added in schools with high levels of transience. For this reason, as well as avoiding the ecological fallacy in estimates of SES on achievement, it would be preferable to have students identified by a unique identifying number as they are in states such as Victoria.

Recommendations:

5. Consideration be given to the introduction of unique student identifying numbers, to improve the precision of like-school and value-added calculations.

4. Precision

Although adoption of unique student identifiers would increase the precision of estimates of student achievement, such estimates of will always include a great deal of statistical uncertainty. One strategy for decreasing uncertainty in student achievement-based estimates of school effectiveness of has been to use more sophisticated statistical models. Multilevel modelling, the preferred technique, structures data at the level of the student, the classroom, the department, the school and the educational system. Attempts to take account of the effect of SES on achievement might, for example, include student level data on parental occupational status, classroom level data on the proportion in professional occupational, and school level data on the average occupational status of the parents in the school (Fitz-Gibbon, 1996, p. 128). Similarly, multilevel modelling can take account of the contribution of other social differences such as gender, home language and cultural group membership. The statistical advantage of this complex form of regression analysis is that the degree of variation attributable to non-random allocation of students to classes or schools can more accurately be calculated.

Despite the theoretical superiority of this approach, it has been argued that there are several reasons to prefer comparisons among schools based on simple regression analysis or multilevel analysis. First, it is argued, the effect of multilevel analyses on small cohorts is to 'shrink' the residual. A small class or school showing very high levels of improvement in raw scores, for example, would have lower residuals when calculated by multilevel than simple regression analysis (Fitz-Gibbon, 1996, p. 130). Interesting results from small cohorts are thus suppressed. Second, it is argued that since the more complex analytical strategy produces residuals which are very similar to those produced by simple regression analysis, it

is preferable to provide schools with results based on a statistical strategy that most teachers understand (Fitz-Gibbon, 1997, p. 44; Tymms, 1999, p.65; Brighton, 2000, p. 130).

In the case of the WALNA data, however, the absence of individually identified data prevents the Data Club from exercising the option of multilevel analysis. For this reason, simple regression strategies will be used.

Recommendation:

6. In the absence of unique student identifying numbers, simple regression rather than multilevel modelling be used to make like-school and value-added calculations.

5. Representation of uncertainty

One of the strongest arguments against the calculation of like-schools and value-added estimates of school effectiveness based on student assessment data is that inexperienced users are likely to overestimate the certainty of the results (Goldstein & Woodhouse, 2000; Myers & Goldstein, 1997). Claims such as these made by Stone (1999,) clearly exceed the statistical certainty of the measurement techniques:

> Value-added assessment's use as an impartial and objective gauge of teacher effectiveness is not its only virtue… value-added assessment can be used to appraise fairly and accurately school and system performance regardless of differences among entering students. In fact, it can be used as an effective indicator in judgments about matters ranging from teacher licensure, tenure, and merit pay to the effectiveness of curricular innovations and teacher training programs.

Better than most, the statisticians who have developed value-added measurement strategies know that their calculations can only be used 'fairly and accurately' within a band of probability. Value-added residuals are based on regression analyses between variables that typically have correlations of about 0.7. Such correlations carry substantial bands of uncertainty within which results have no statistical significance. In the hands of managers and journalist who wish to make data 'user friendly', technical language and statistical uncertainty may be eliminated. Figure 9, for example, shows a table from the *Tennessean* identifying a set of schools as 'best' and 'worst' by adding together the Tennessee value-added assessment system's six raw score letter grades and the six value-added letter grades. Two different kinds of data are conflated, and there is no explanation of the statistical uncertainty behind these grades.

Unhelpful and over-determined secondary analyses such as these have lead some authorities such as Rowe (2000, p. 85) to argue that the presentation of uncertainty intervals should be as prominent in the data display as any performance indicator values such as value-added residuals. The alternative argument is that publication of uncertainty intervals or standard errors is an attempt to indicate whether the result is the consequence of a sampling error. The consequence of this strategy is that many schools will show no statistically significant value added. In a series of

comparisons between methods of estimating value-added, CTF found that between 11-18 percent of schools were associated with lower than expected value-added and 12-18 percent were associated with higher than expected value-added scores. That is, a total of between 77 and 64 percent showed no significant difference from the predicted level of value added (Fitz-Gibbon, 1997, p. 39). Similar results have been reported by Rowe (2000) and Goldstein (2001) and Saunders (1999). That is, value added measures 'may be able to establish that differences exist among schools … they cannot, with precision, indicate how well a particular school is performing' (Rowe, 2000, p. 81). Moreover, as Fitz-Gibbon has reported, because school value-added scores vary from year to year (1997, p. 41) a trend several years performance may be useful that a single year residual calculation.

Figure 9. Simplified reporting of value added calculations

Whether results are doe to a sampling error or not, argues Brighton (2000, p.130), the calculated value is the best estimate available until the next time the data are measured. Sometimes even small positive or negative residuals with the band of statistical uncertainty will be of interest to schools in considering the effectiveness of their programs. Provided that the data display includes some measure of the uncertainty inherent in production of the results, schools may exercise an informed choice about whether to attend to apparent differences that lie within the band of statistical uncertainty. The British experience with value-added residuals has been that schools participating in a voluntary value-added feedback

program find the information useful and understand the need not to over-interpret the results (Goldstein, Huiqi, Rath, & Hill, 2000).

Recommendation:

7. Statistical uncertainty as well as calculated school values be represented in like-school and value-added data displays a.

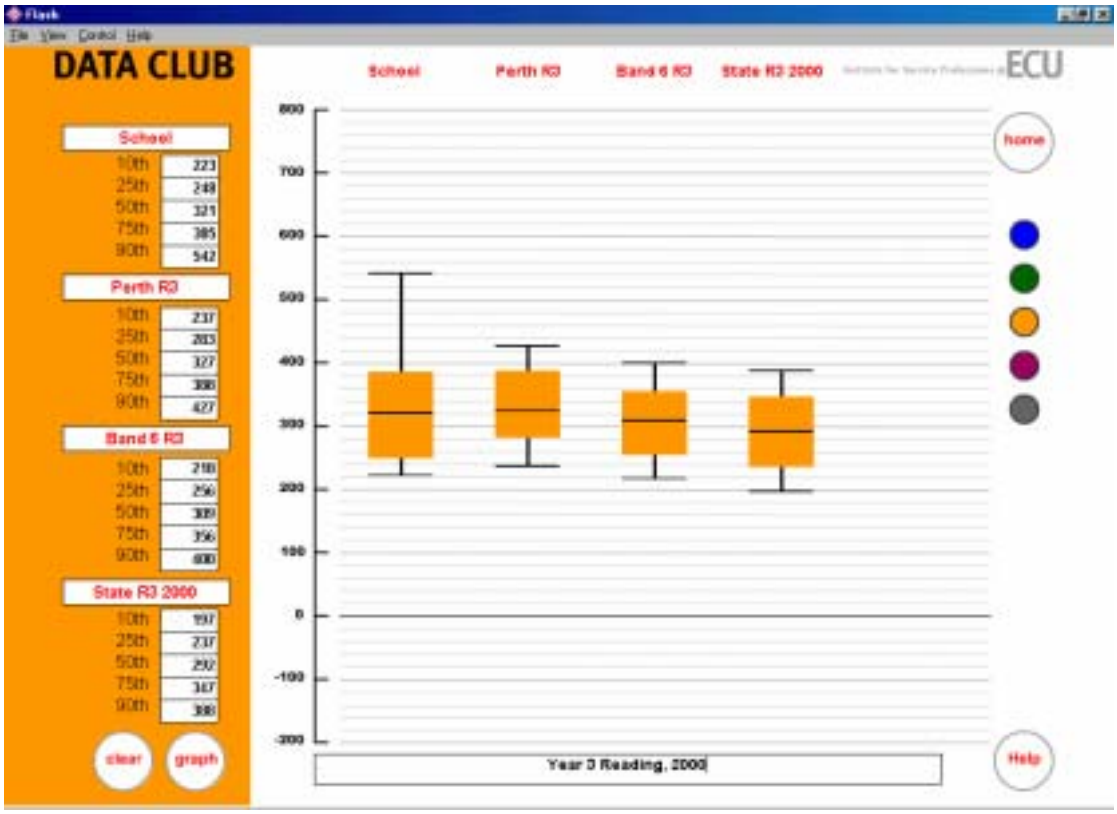
DATA DISPLAY

Round 2 of the Data Club provides three forms of analysis, developed to be consistent with recommendations of the literature review (above).

1. Student performance, 2000

This first level analysis provides student data in percentiles and uses a box and whisker plot to provide a visual comparison between the distributions for the school, a group of like schools based on the H index, the school district and the State. There are nine H index bands, numbered from 0 to 8 in increasing SES order. Data plotted are 2000 WALNA scores in reading, writing, spelling and numeracy. No statistical inferences are drawn about similarities or differences among the distributions. A sample data display appears in Figure 10. The custom-designed graphing program used in the display will be available to schools in September 2001 at http://isp.ecu.edu.au. The state, district and like-school data are stored on the downloadable program and the confidential school data are entered by schools from the Data Club confidential school data sheets.

Figure 10. Box and whisker plots of school, like school, district and State scores

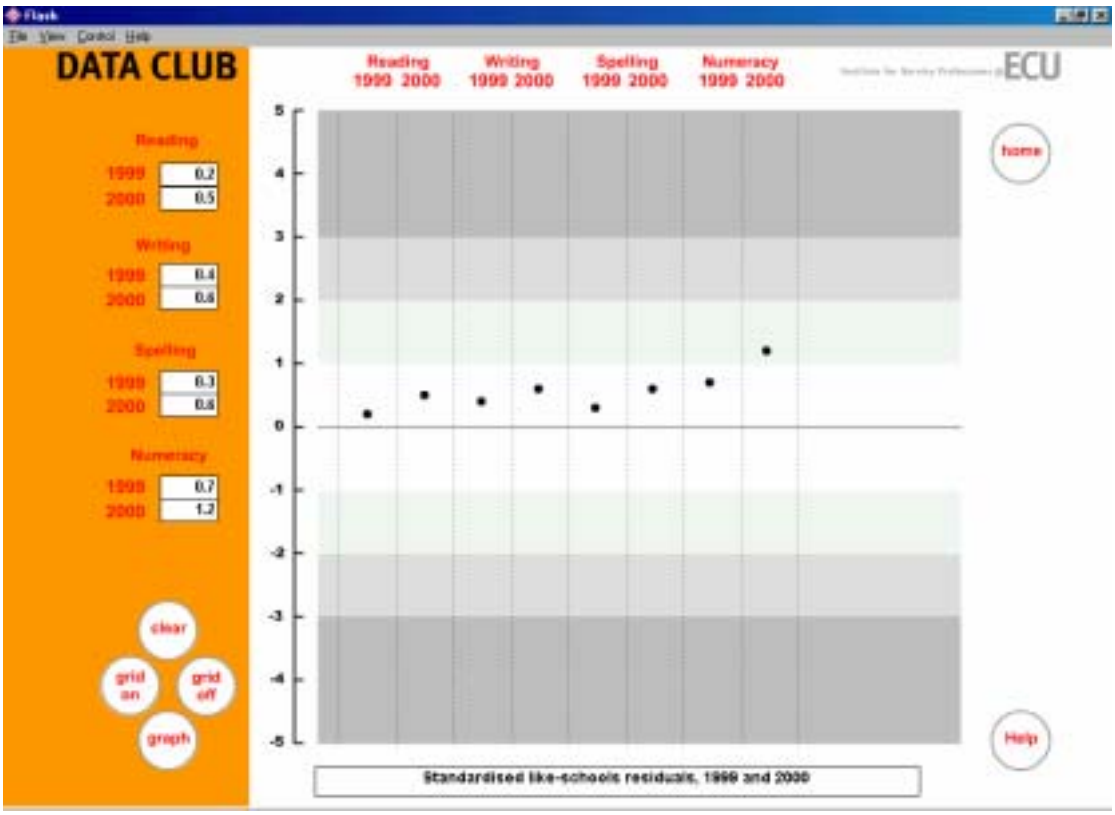


2. Like schools, 1999 and 2000

This analysis attempts to answer the question 'How well did the school perform in 1999 and 2000 compared with similar schools?' Based on a simple regression analysis that uses school H index to predict school mean score, this analysis reports reading, writing, spelling and numeracy scores in 1999 and 2000. School scores are reported in standardised residuals. The data display uses a convention of light to dark shading to represent degrees of statistical certainty. Scores in the range ±1 have a probability of up to 68 percent that the school's real score is not different from zero. These scores appear in the central white zone on Figure 11. Scores in the range +1 to +2 and -1 to -2 have a probability of up to 95 percent that the school's real score is not zero. These scores appear in the lightest grey zones. Scores in the range +2 to +3 and -2 to -3 have a probability of up to 99.9 percent that the school's real score is not zero. These scores appear in the mid grey zones. Scores in the range beyond ±3 have a probability of greater than 99.9 percent that the school's real score is not zero. These are represented in the darkest grey zones.

> The data display strategy here is to use the shaded background to represent the probability that the result is not due to chance, but also to allow the possibility that small and not statistically significant results still may be of interest to schools. In Figure 11 for example, although only numeracy 2000 has a value approaching statistical significance, the greater residual for all 2000 scores than the corresponding 1999 residual may suggest to schools that the 2000 cohort is stronger than the 1999 cohort, other things being equal.

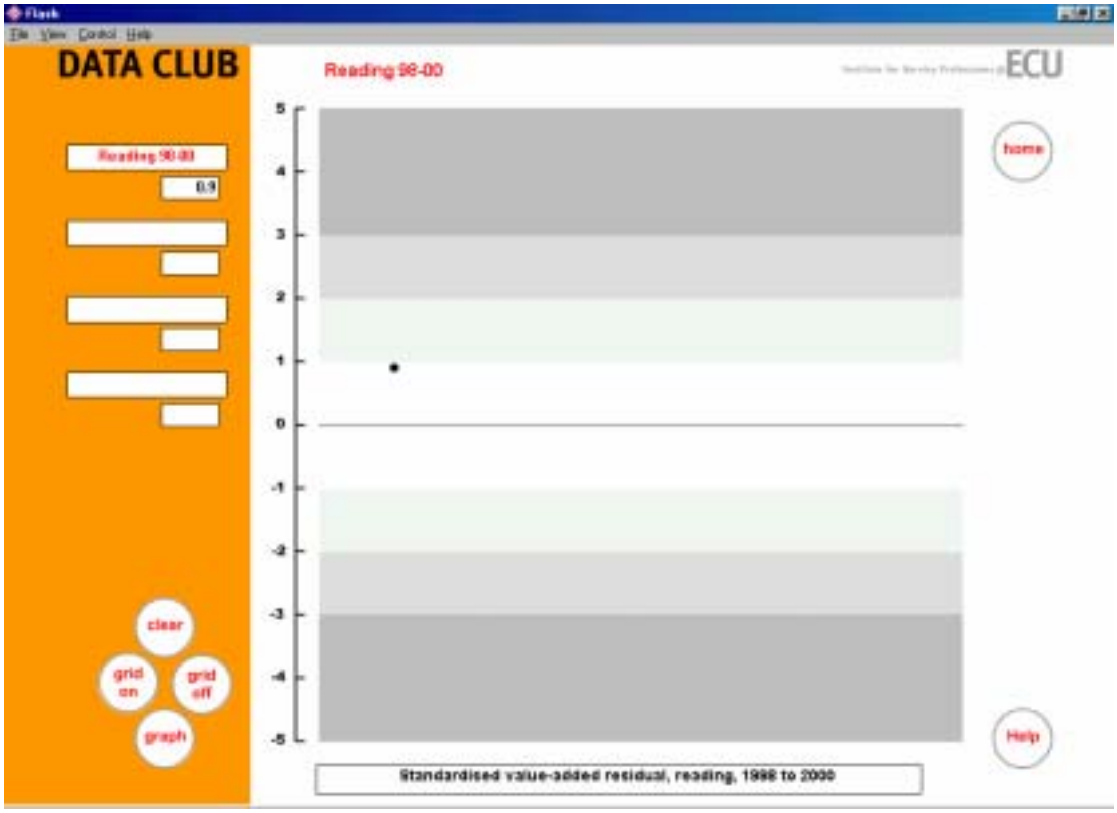
Figure 11. Standardised like-school residuals, 1999 and 2000



3. Value-added, 1998 to 2000

This analysis attempts to answer the question 'How much value is this school adding over time in reading?' As the WALNA data set currently includes only one set of prior and subsequent achievement scores, value-added from Year 3 in 1998 to Year 5 in 2000 can be calculated only in reading. In 2002, value added residuals will be calculated for growth from Year 3 (1999) to Year 5 (2001) and Year 5 (1999) to Year 7 (2001) in all assessment domains. The procedure used is a simple regression analysis that uses the school mean score for reading in 1998 to predict the school mean score in reading in 2000. The same data display uses a convention of light to dark shading to represent degrees of statistical certainty. As was the case in the like-schools residual analysis, the data display strategy is to use the shaded background to represent the probability that the result is not due to chance, but also to allow the possibility that small and statistically insignificant results may be of interest to schools. In this case, the standardised residual of +0.9 indicates that the Year 5 (2000) reading scores are higher than predicted from the Year 3 (1998) reading scores. Although the residual does not reach the usual statistical significance criterion of 95 percent confidence (±2) the residual of +0.9 provides the school with no cause for concern about student progress relative to the other schools in the state.

Figure 12. Value added residual, reading 1998 to 2000



REFERENCES

Ainley, J., Graetz, B., Long, M., & Batten, M. (1995). *Socioeconomic status and school* education. Canberra: AGPS.

Board of Studies (Victoria). (2000). *AIM Testing 2000 Reporting Guide*, Carlton, Vic.: Author.

Brighton, M. (2000). Making our measurements count. *Evaluation and Research in Education, 14*(3&4), 124-35.

Farish, S. (1993). *Constructing census-based indicators of (educational) disadvantage: A summary*. Mimeo.

Fitz-Gibbon, C.T. (1996). *Monitoring education: Indicators, quality and effectiveness*. London: Cassell.

Fitz-Gibbon, C.T. (1997). *The value added national project. Final report. Feasibility studies for a national system of value-added indicators.* (http://www.cem.dur.ac.uk/)

Goldstein (2001). *Using pupil performance data for judging schools and teachers: Scope and limitations*. London Institute of Education.

Goldstein, H., & Woodhouse, G. (2000). School effectiveness research and educational policy. *Oxford Review of Education*.

Goldstein, H., Huiqi, P., Rath, T., & Hill, N. (2000). *The use of value added information in judging school performance*. London: Institute of Education.

Hill, S., Comber, B., Louden, W., Rivalland, J., & Reid, J. (2001). *100 children turn* 10. (Volumes 1-2). Canberra: Department of Education, Employment, Training and Youth Affairs.

Marks, G.N., McMillan, J., Jones, F.L., & Ainley, J. (2000). *The measurement of soccioeconomic status for reporting of nationally comparable outcomes of schooling.* Draft Report for the national Education Performance Monitoring Taskforce. Melbourne: ACER & Canberra: Sociology Program, Research School of the Social Sciences, ANU.

Masters, G., & Forster, M. (1997). *Literacy Standards in Australia*. Camberwell, Vic.:ACER.

Myers, K., & Goldstein, H. (1997). Failing schools in a failing system. In A. Hargreaves (Ed). *Rethinking educational change with heart and mind.* ASCD.

O'Donoghue, C., Thomas, S., Goldstein, H., & Knight, T. (1997). *The 1996 DfEE study of value added for 16-18 year olds in England.*. London: Institute of Education.

Rogosa, D. (2000). *Interpretive notes for the Academic Performance Index*. California Department of Education. (http://www.cde.ca.gov/psaa/api/fallapi/apnotes.pdf)

Ross, K.N., Farish, S., & Plunkett, M. (1988). *Indicators of socio-economic disadvantage for Australian schools*. Geelong: Deakin Institute for Studies in Education.

Rowe, K.J. (2000). Assessment, league tables and school effectiveness: Consider the issues and 'Let's get real'! *Journal of Educational Enquiry, 1*(1), 73-98.

Sanders, W.L. (1998). Value-added assessment. *School Administrator, 55*(11), 24-27.

Sanders, W.L., & Horn, S. (1994). The Tennessee value-added assessment system (TVASS): Mixed model methodology in educational assessment. *Journal of Personnel Evaluation in Education, 8*, 299-311.

Stone, J.E. (1999). Value-added assessment: An accountability revolution. In M. Kanstoroom & C.E. Finn, Jr., *Better teachers, better schools.* Washington DC: The Fordham Foundation. (http://www.edexcellence.net)

Teese, R. (2000). *Academic success and social power: Examinations and inequality.* Melbourne: Melbourne University Press.

Thomas, S., & Mortimore, P. (1996). Comparison of value-added models for secondary school effectiveness. *Research Papers in Education, 11*(1), 5-33.

Tymms, P. (1999). *Baseline assessment and monitoring in primary schools: Achievements, attitudes and value-added indicators*. London: David Fulton.

Western Australia, Department of Educational Services (Robson Report). (2001). *Investing in government schools: Putting children fist. The report of the taskforce on structures, services and resources supporting government schools.* Perth, W.A.: Author.

Western Australia, Education Department. (1995). *Monitoring standards in education '94 achievement. Overview***.** Perth, W.A.: Author.

Western Australia, Education Department. (1999). *Student achievement in mathematics in Western Australian Government schools, 1998*. Perth, W.A.: Author.

Western Australia, Ministry of Education. (1991). *Educational standards in Western Australian government schools 1990.* Perth, W.A: Author.

Yang, M., Goldstein, H., Rath, T, & Hill, N. (1999). The use of assessment data for school improvement purposes. *Oxford Review of Education, 25*(4), 469-83.

Appendix 3: Web-site Report