



OXFORD UNIVERSITY CENTRE FOR EDUCATIONAL ASSESSMENT

The following paper was presented at the 34th International Association for Educational Assessment Conference at Cambridge UK, 9th September 2008.

Referenced as:

Stanley, G & Tognolini, J. (2008) Performance with respect to standards in public examinations. 34th IAEA Conference, Cambridge, UK.

PERFORMANCE WITH RESPECT TO STANDARDS IN PUBLIC EXAMINATIONS

Gordon Stanley & Jim Tognolini

Oxford University Centre for Educational Assessment

Abstract

Public examination results are scrutinized by the media and the public each year with respect to whether or not 'standards' are rising or falling. From a technical point of view the debate which ensues is about the numbers attaining or not attaining a particular grade or bench mark. These grades or benchmarks represent the achievement standard. Hence 'standards' should not be considered to be changing but the numbers reported with respect to the standards can change. The debates centre around the extent to which reported changes in numbers achieving the standard are credible and represent 'real' changes in performance of students or simply changes due to the examination and reporting process. Most public examination systems which use a standards-referenced system of reporting report some incremental creep. This paper examines some similarities and differences across subject areas and systems.

Results from public examinations in senior secondary schooling are used for competitive selection purposes ranging from university entrance and scholarships through to employment. Given the use to which the results are put, they can be considered 'high-stakes' examinations. Senior secondary certificates of education typically report subject performance in terms of grades or standards of performance.

In most countries examination authorities face media scrutiny each year with the release of results. Commonly there is debate about whether or not 'standards' are rising or falling. The trigger for the media debate is any variation in the proportion of students attaining or not attaining a particular grade or benchmark. From a technical point of view 'standards' should not be regarded as changing, but technical niceties do not make for juicy headlines.

The media problem is caused by the move away from normative equating procedures for reporting results. Inevitably in every education system with high-stakes assessment there is strong competition in attaining the highest grade. When results are normalised or fitted to a normal curve it is relatively easy to have a fixed proportion of candidates achieving the highest reported marks each year. Such systems typically report 4-6% in their highest-grade level (Sadler, 2005,p186). When normative scaling is applied to all subjects the percentage reported as achieving the top grade in each subject is essentially the same. In such systems the reporting preserves the ranking of student performance but does not provide information about the content of the achievement. However the virtue of contrived consistency of results is contrary to modern reporting requirements (Tognolini & Stanley, 2007).

The outcomes focus of modern education systems has resulted in a move away from a statistical equating of results towards a standards-setting model based on achievement of specified performance standards. In such an environment there is less control by the examination authority of the numbers achieving the highest grade within and between subjects. The characteristics for recognition of high performance in a standards model are typically spelled out in grade descriptions which are used to identify exemplars which define the achievement. For assignment of grades to occur judgments are made about whether or not the appropriate standards have been demonstrated.

One of the problems facing systems reporting with respect to standards is the meaning attached to variation in the numbers achieving the top grade over time. Time series data often show incremental creep with more students achieving the top levels of performance each year. This result then leads to debate about whether or not standards are falling or whether the education system itself is delivering some consistent improvement (Wikstrom, 2005).

Two potential sources of difference can occur in a standards model of reporting. First differences can occur between subjects at the level of standards setting. Even when the same generic performance descriptors are used their application across subjects can result in different levels of difficulty: some subject standards may be harder to achieve than others. Certainly there is a long entrenched view about

the relative toughness of different academic disciplines (see Bourdieu,1988), which makes equating of performance standards drawn from different subject curriculum content standards somewhat difficult.

Secondly, differences between systems in the numbers reported achieving the highest grade in the same subject could be due to differences in the standards-setting process used. There are a number of different standards setting processes employed by education authorities that manage public examination systems. While there are a range of views about the merits of different standards-setting procedures it has been found that outcomes are influenced by the procedure adopted as well as the standards adopted (Cizek, 2001). When bench-marking performance across education systems these differences in procedure need to be considered as well as any differences in the content of standards adopted by the education authority.

In an era of concern about comparative performance there has been little comparative analysis of the similarities and differences in reporting outcomes across subjects between different education systems when a standards-setting process is used. This paper compares top grade performance data for ten subjects reported by two assessment authorities in the United Kingdom (the British Joint Council for Qualifications and the Scottish Qualifications Authority) and two from Australia (the Queensland Studies Authority and the Board of Studies, New South Wales).The UK and Queensland authorities have had a standards-based grade reporting system for some years. In NSW the Board of Studies changed from norm-referenced scaling of all subjects to standards-referenced reporting in 2001.

For the purpose of the current study the following ten traditionally academic subjects assessed by each of the four qualifications authorities were selected: English, French, German, Mathematics, Biology, Chemistry, Physics, Economics, Geography and History. Candidature size across these subjects were such that one would expect results to be less subject to effects due to cohort differences from year to year than would be expected in courses with small enrolments.

Making judgements about the comparability of the curriculum in these four systems is difficult given the different ways in which content may be specified in official documents, and implemented in the classroom. Moreover there may be significantly different drivers of subject choice across systems. Nevertheless for traditional academic subjects it is assumed that, even when local differences in curriculum are acknowledged, there is considerable common intellectual content across education systems.

METHOD

Results data from 2001-4 for the ten subjects were obtained from the British Joint Council for Qualifications (JCQ)for A Level GCE results (sourced from <http://www.jcq.org.uk>), from the Scottish Qualifications Authority (SQA) for their New Higher Grades (sourced from <http://www.sqa.org.uk>), from the Queensland Studies Authority (QLD) for their Senior Secondary Certificate (sourced from

<http://www.qsa.qld.edu.au>) and from the Board of Studies New South Wales (NSW) for their Higher School Certificate (sourced from <http://www.boardofstudies.nsw.edu.au>).

Three of the four authorities have public examinations while the Queensland Studies Authority uses moderated school assessment of student portfolios to arrive at grades. The UK systems use a standards-setting process, which involves consideration of performance data as well as statistical data. In NSW a modified Angoff standard-setting procedure is used without the judges knowing the distributional consequences of their cut-score decisions (see MacCann, & Stanley, 2004).

RESULTS

The A Level GCE results are reported on a five level scale from E to A; the New Higher results from SQA are reported on a four level scale from Pass, C, B to A; the QLD report on a five level scale from VLA to VHA and NSW report on a six level scale from band 1 to 6. For the purpose of the present report the percentage achieving the highest grade reported (A, VHA or Band 6) was compared.

The education systems differ in the number of grades reported as well as in the number of subjects taken by students. While students in England typically take three A-levels, for the SQA, QLD and NSW authorities five subjects are usually taken.

Across the years 2001-07 the average percentage of students in the top grade for the four systems are presented in Figure 1. Apart from French and German, the UK systems tend to have on average about 10% more students achieving their top grade than in the Australian systems.

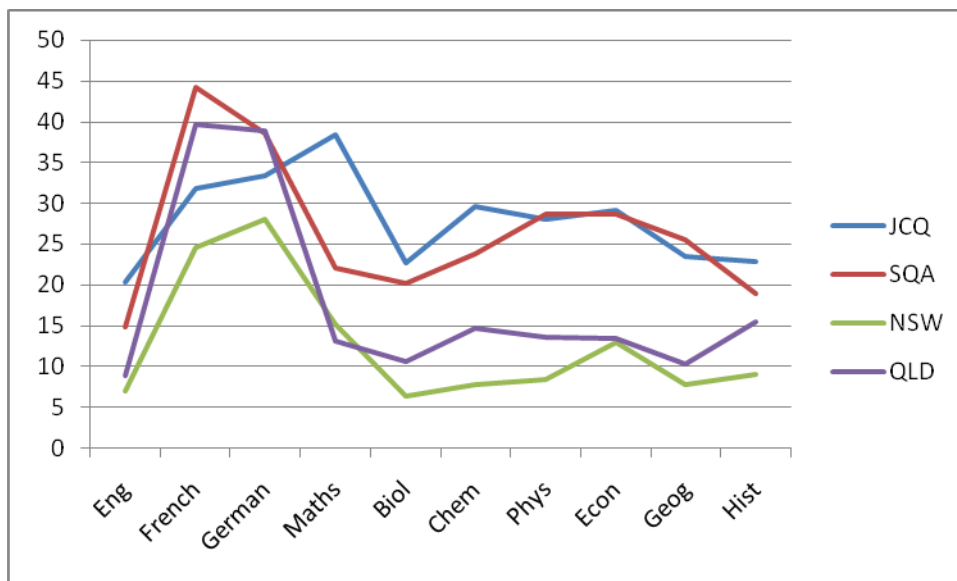


Figure 1: Average percentage of students in top grade for each authority across 10 subjects

A common pattern across systems is for English to have the lowest percentage, for French and German to have the highest, for Biology to be lower than the physical sciences and for Economics to be higher than Geography and History. These trends presumably reflect some common aspects of student selection or relative subject standards across the systems.

Table 1 shows the means and standard deviations for each subject for each authority.

	Means				Standard Deviations			
	JCQ	SQA	NSW	QLD	JCQ	SQA	NSW	QLD
English	20.26	14.86	6.97	8.83	2.15	1.86	1.53	0.51
French	31.81	44.14	24.58	39.68	3.86	2.12	3.20	1.68
German	33.36	38.57	28.00	38.82	3.49	1.72	2.77	1.90
Maths	38.40	22.14	15.07	13.09	4.96	1.95	2.00	1.45
Biology	22.56	20.14	6.36	10.63	2.36	4.22	2.86	0.69
Chemistry	29.57	23.86	7.84	8.46	1.88	4.45	2.18	1.69
Physics	28.01	28.71	8.46	13.60	1.92	1.89	2.70	2.25
Economics	29.06	28.57	12.93	13.46	3.83	3.15	1.53	1.78
Geography	23.34	25.43	7.75	10.24	2.87	1.81	2.89	1.29
History	22.80	19.00	9.07	15.45	2.45	2.45	1.20	1.08

Table 1: Means and standard deviations for percentage of top grade in subjects at each authority averaged from 2001-2007.

From this table it can be seen that as well as differences across subjects there are differences in the amount of variability of these means across subjects and across systems. The linear trends over time for each of the subjects are shown in Figures 2-11.

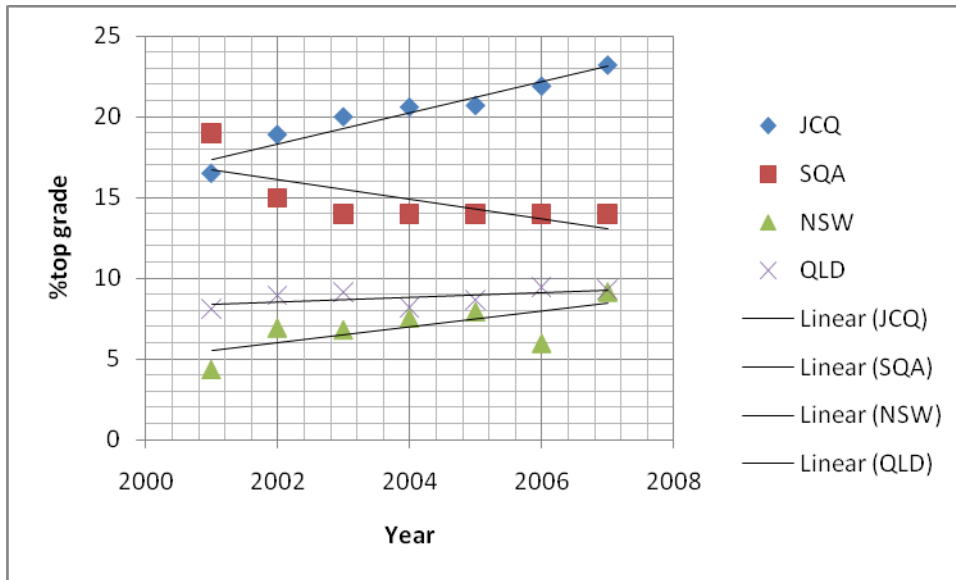


Figure 2: Trend for English top percentage

Of interest in Figure 2, which shows the trends for English, is the divergence over time between the results for JCQ and SQA, while the Australian trends are converging.

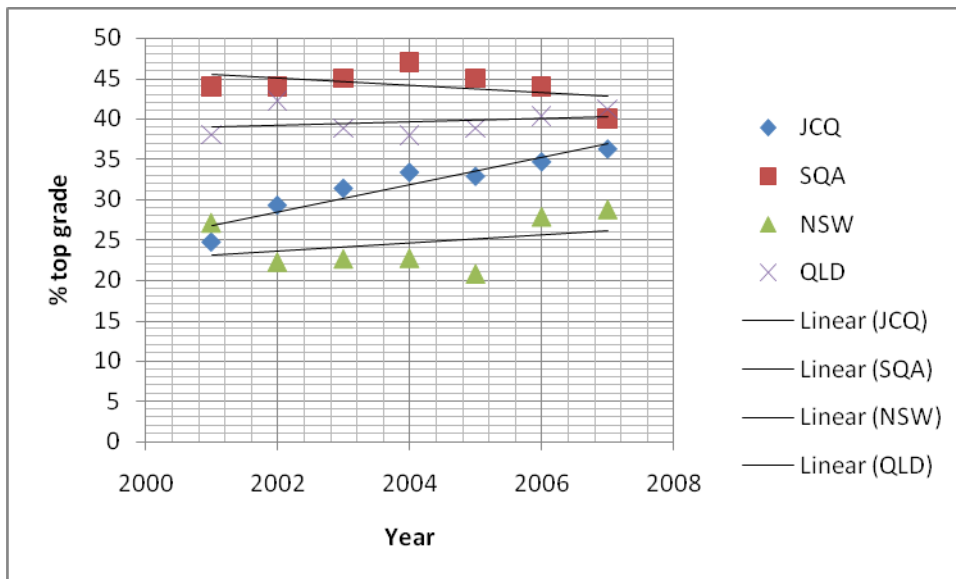


Figure 3: Trend for French top percentage

In Figure 3 which presents the comparison for French we can see that two authorities have a positive trend while QLD is relatively stable and SQA has a small decline.

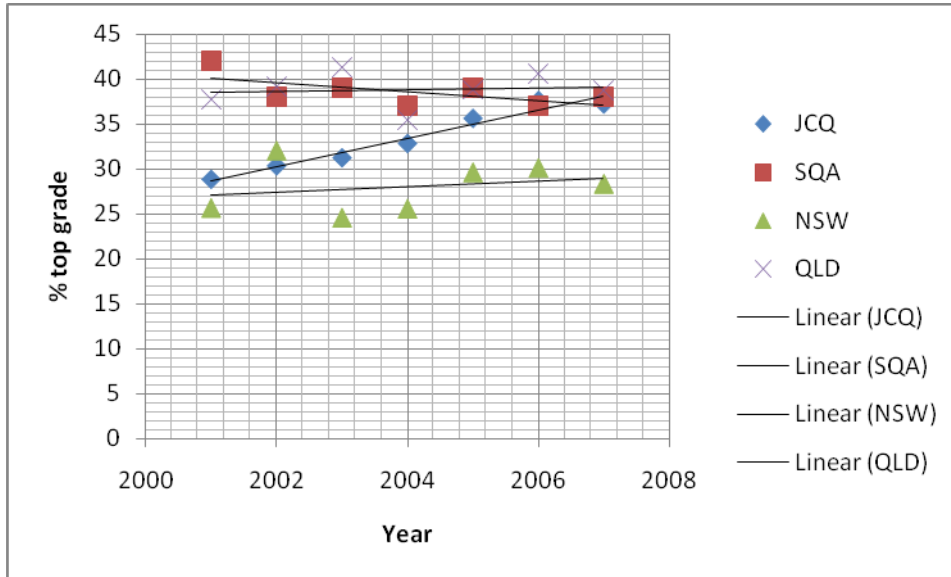


Figure 4: Trend for German top percentage

In Figure 4 German has a similar trend pattern over time across authorities as French.

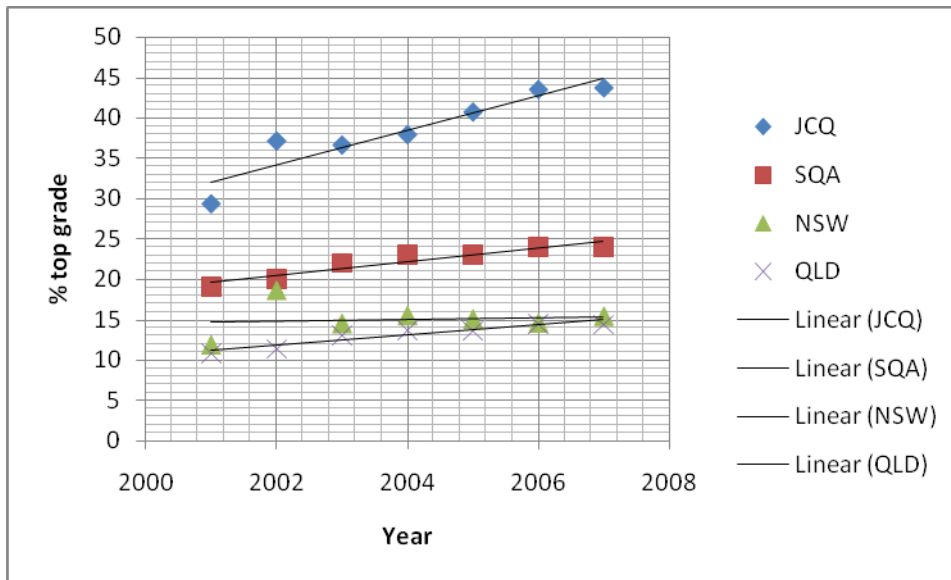


Figure 5: Trend for Maths top percentage

Apart from NSW, the other three authorities all manifest an upwards trend over time for top grade in Maths.

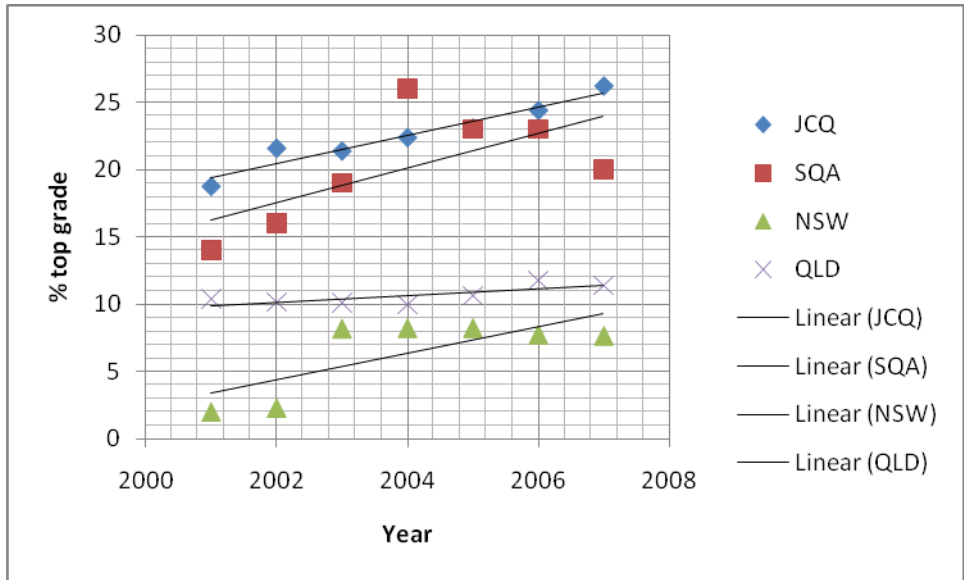


Figure 6: Top percentage for Biology

As shown in Figure 6 in Biology the upward trends show some variability from a linear fit for both SQA and NSW.

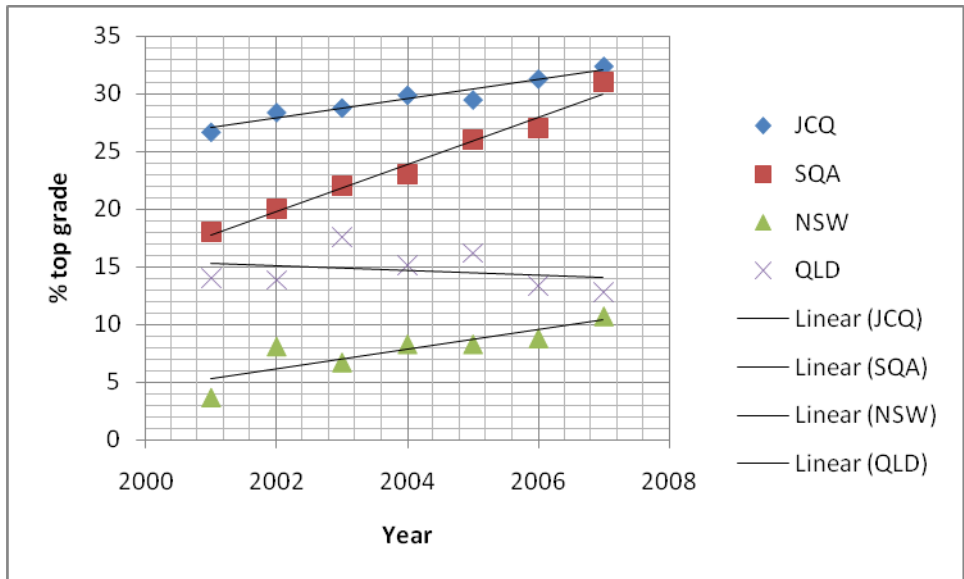


Figure 7: Top percentage for Chemistry

With Chemistry a positive trend over time occurs for three authorities with QLD showing a relatively stable outcome over time.

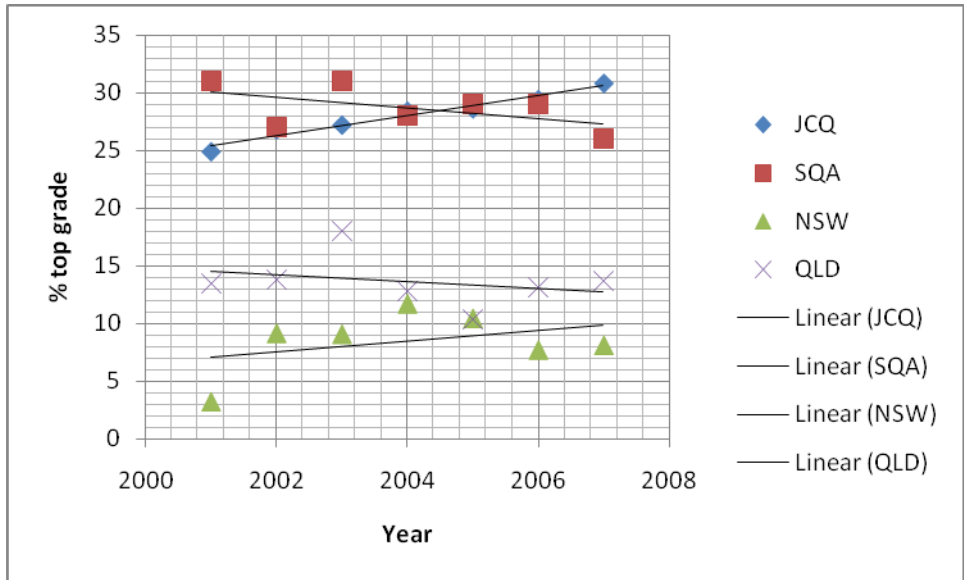


Figure 8: Top percentage for Physics

In Figure 8 we can observe that for Physics both SQA and QLD show a downward trend while JCQ and NSW show an upward trend.

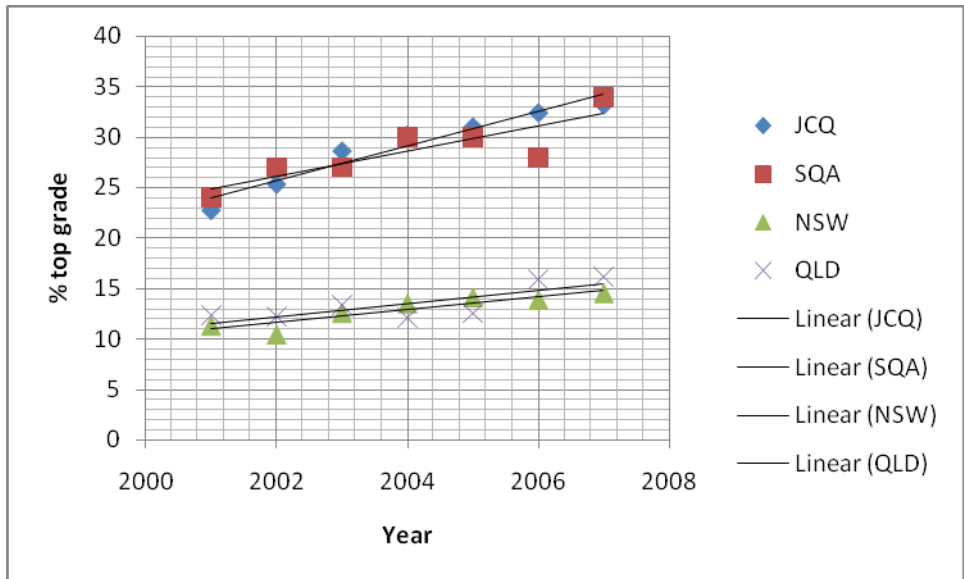


Figure 9: Top percentage for Economics

The trend in Figure 9 for Economics is interesting in showing the closeness of trend for the two UK authorities and the closeness for the Australian authorities. For both countries there is an upward trend.

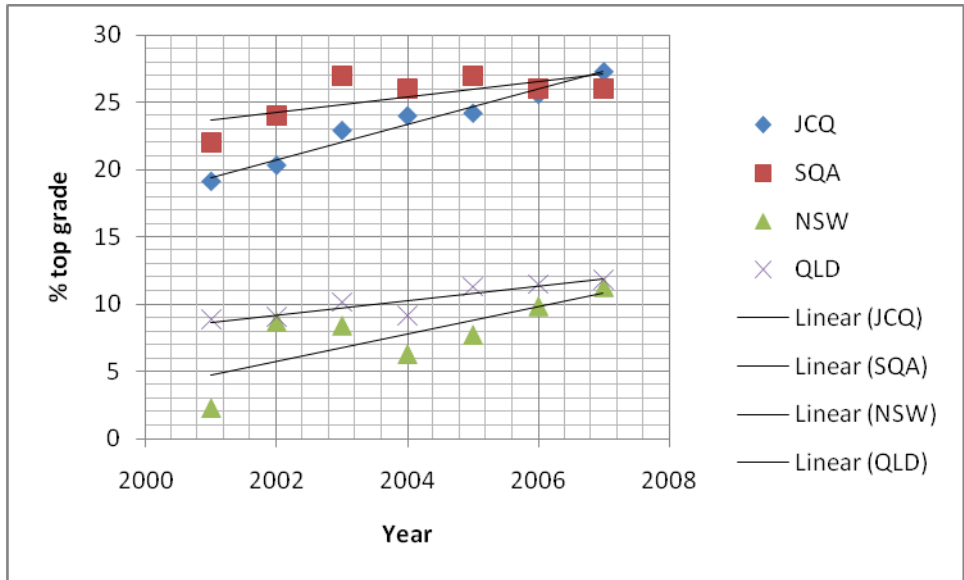


Figure 10: Top percentage for Geography

Figure 10 shows incremental creep for Geography over time for all systems with convergence for the two authorities in each country.

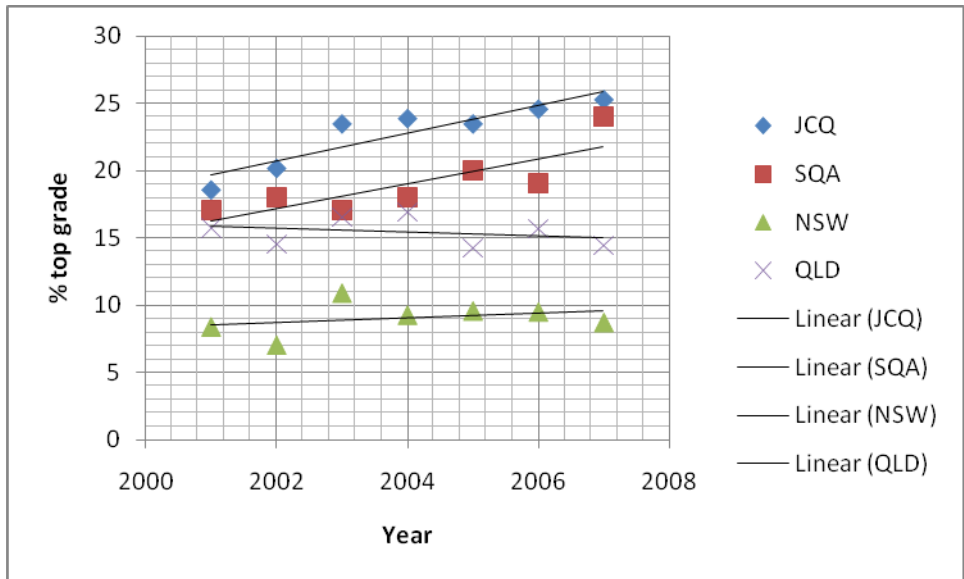


Figure 11: Top percentage for History

The pattern for History shown in Figure 11 indicates incremental creep for both JCQ and SQA and relative stability for QLD and NSW.

In figures 2-11 it can be seen that JCQ has incremental year-on-year creep for all subjects, while incremental creep does not occur across all subjects in the data from the other authorities. For other

authorities the patterns differ across subjects and authorities as to whether or not there is incremental creep, stability, or a downward trend. However incremental creep is a more common trend than stability or a downward trend.

DISCUSSION

Comparing the four systems shows some consistency in relative differences in the magnitude between the top grade performances across subjects. However the trend towards upward creep over time shows different patterns across systems with respect to subjects. Only JCQ has consistent creep for all subjects. Other systems have it occur in some subjects but not others.

The consistency across all subjects selected for analysis of incremental creep in the top grade English A levels is of considerable interest. While consistent improvement over time due to better pedagogy is possible it is highly unlikely that England is more successful in achieving a consistent improvement across subjects than Scotland. Today all education systems are under similar pressures to demonstrate improvements in student performance. It would be comforting to think that incremental creep was primarily due to 'real' improvement in subjects by students in the education system. Nevertheless at present we cannot be confident that particular features of the standards setting process are not primarily responsible for the differences in reported outcomes

Having a relatively high percentage achieving the highest grade can lead to argument that the standard is set too low and that there is not enough challenge for the more talented students. Clearly whether or not this is a valid concern for qualifications authorities will depend on the needs of their system. At approximately 25% on average the UK systems have settled on a higher percentage achieving their top grade than is the case for the Australian systems, which typically report in the 10-15% range. This result may be influenced by the difference in significance of the top grade for university entrance. In the Australian systems subject performance is scaled statistically to produce a university entrance rank, so the subject achievement level is less prominent in the selection process than in the UK.

As mentioned earlier the average percentage for the top grade may be due in part to the specific standards-setting procedure adopted by the authority. Different standards-setting procedures can have some effect on the numbers reported achieving the highest level. Green, Trimble and Lewis (2003) reported differences between three standards-setting procedures used to set cut scores in each of 18 grade/content areas in the Kentucky state assessment system. Their results showed method difference of about 8% from the lowest to highest cut for the top level and this was relatively consistent for each method across subjects.

Bench-marking and equating standards across systems is difficult because of differences in curriculum and assessment procedures. Judgements of performance with respect to standards as well as definitions

of the standards themselves are contextually determined. Despite all the differences, which should work against similarity, the present study has shown that there is some consistency in the relative pattern of numbers achieving the top grade in particular courses across systems.

Presumably the pattern reflects some common features of the differences between academic disciplines. While grade descriptors for high achievement tend to have a semantic similarity stressing excellence and complex reasoning they require different subject content to be mastered by students. Despite valiant attempts by curriculum writers to equate difficulty of content across subjects, it is hard to achieve in practice. An example of the descriptors for Economics and French for QLD and NSW are presented in Table 2. From this table it appears easier to interpret similarity within the subject discipline than it is across the subject disciplines.

Where there is choice of subject it may well be the case that there are differences in the ability level of students who choose particular subjects and this tendency is relatively consistent across education systems. For example, the higher number of students achieving the top grade in French and German may be partly due to weaker language students dropping out when the assessment is high stakes. An alternative possibility is that despite attempts to equate standards across disciplines, the highest standards for languages are somewhat easier than the highest standards in other subjects, though this is not immediately clear from the descriptions in Table 2.

Economics Grade/Band Descriptors for QLD and NSW

QLD VH A - Has accurate and comprehensive knowledge, understanding and recall of facts, concepts, contexts, principles, underlying theories and econometric models from the course. Analyses and organises information in a comprehensive manner Accurately comprehends economic information in a variety of contexts.

Consistently accurate in analysis of trends, patterns and cause-effect relationships. Applies learnt knowledge and skills in a wide variety of unfamiliar situations. Independently draws on information from a wide range of sources and combines them into a coherent whole. Develops and uses a range of appropriate criteria to evaluate alternative ideas, proposals or solutions to economic problems. Adapts and manipulates the inquiry process to reach decisions about proposals, issues and hypotheses. Independently gathers, records and checks detailed information from a variety of sources including primary sources. Critically selects relevant data and information and structures them to achieve defined purposes and outcomes within a specified time. Uses mathematical techniques and language and referencing conventions accurately. Ideas and information have been communicated concisely in a variety of genre and forms appropriate to context.

NSW Band 6 - Integrates economic terms, concepts, relationships and theory in a variety of economic contexts. Displays superior analysis of the role of economic participants and markets in a variety of economic contexts. Uses extensive economic vocabulary and illustrative examples in exposition of problems and policies in a variety of contexts. Demonstrates critical judgment and sound reasoning to select, organise, synthesise and evaluate relevant information from a variety of sources. Presents excellent explanation and evaluation of the impact of government economic policies in contemporary and hypothetical economic contexts. Presents comprehensive application of appropriate mathematical concepts in a variety of economic contexts. Produces comprehensive economic arguments to evaluate the consequences of economic problems and issues on economic participants.

French Grade/Band Descriptors for QLD and NSW

QLD VHA - The student... conveys meaning clearly, uses a wide range of vocabulary & structures, displays flexibility in sentence structure, uses a range of complex sentences which may include aspects of time, mood & intention, shows some originality. Familiar language (including spelling, punctuation & word order) is mostly accurate. Communication is clear although errors may occur in more complex language. Register is appropriate. Work is relevant to task. Work is... well organised, cohere, relevant in content, length & format. The student... shows a comprehensive understanding of main idea, distinguishes main points from minor points, gist from detail, deduces meaning from context, draws appropriate conclusions, infers speaker's intentions & attitudes, recognises register. The student... conveys meaning clearly, some errors may occur, shows some awareness of sociocultural elements, conveys intention & attitude successfully, initiates & sustains a conversation, develops ideas coherently, usually uses appropriate pause fillers & non verbal techniques when required. Features are acceptable to a sympathetic background speaker.

The student...shows a comprehensive understanding of main ideas, distinguishes main points from minor ones, gist from detail, deduces meaning from context, draws appropriate conclusions, infers purpose of text and attitude of writer, understands common socio-cultural references, recognises tone.

NSW Band 6 - Initiates and sustains conversation through the exchange of relevant information and ideas appropriate to context, audience and purpose. Demonstrates a sophisticated command of a wide range of vocabulary and language structures. Manipulates language structures in a creative, authentic and fluent manner, with minor errors. Structures and sequences ideas and information effectively and creatively. Demonstrates a comprehensive global and detailed understanding of French by analysing, processing and responding to spoken and written texts.

Table 2: Economics and French Subject Descriptors for Top Grade for QLD and NSW

The comparison across education systems suggests that whatever factor is at work there is some similarity of outcome when results of students are not statistically equated across subjects. However, while there is no agreement or common practice about how to ensure grade-setting processes are stable with respect to standards, it is difficult to attach educational meaning to changes in the differences in proportion achieving the top grade across subjects or years.

REFERENCES

Bourdieu, P. (1988) *Homo academicus*. Cambridge, U.K.: Polity Press

Cizek, G.J. Ed. (2001) *Setting performance standards: concepts, methods and perspectives*. Mahwah, N.J.: Erlbaum.

Green, D.R., Trimble, C. & Lewis, D.M. (2003). Interpreting the results of three different standards-setting procedures. *Educational Measurement: Issues and Practices*, 22, 1, 22-32.

MacCann, R.G. & Stanley, G. (2004). Estimating the standard error of the judging in a modified-Angoff standards setting procedure. *Practical Assessment, Research & Evaluation*, 9(5): <http://pareonline.net/>

Masters, G.N. (2002). *Fair and meaningful measures?: a review of examination procedures in the NSW Higher School Certificate*. Camberwell, Victoria: ACER.

Sadler, D.R. (2005) Interpretations of criteria-based assessment and grading in higher education. *Assessment & Evaluation in Higher Education*, 30, #2, 175-194.

Tognolini, J. & Stanley, G. (2007). Standards-based assessment: a tool and means to the development of human capital and capacity building in education. *Australian Journal of Education*, 51, 2, 129-145.

Wikstrom, C. (2005) Grade stability in a criterion-referenced grading system: the Swedish example. *Assessment in Education*, 12, 2, 125-144.

About the Authors

Gordon Stanley is Pearson Professor of Educational Assessment and Director of the Oxford University Centre for Educational Assessment.

Jim Tognolini is Senior Research Fellow at the Oxford University Centre for Educational Assessment, and Director Pearson Research and Assessment.

Address

15 Norham Gardens, Oxford
OX2 6PY, UK

Email

gordon.stanley@education.ox.ac.uk

jim.tognolini@pearson.com

Descriptors: high stakes tests, standards-setting, international assessment comparison.

