ANSWERS ON NOTICE 1: Page 13 of the transcript

The CHAIR: Would you like to make closing comments? We did not have any questions taken on notice. However, Mr Tognolini, I was going to ask: In your opening remarks you mentioned research studies showing that so-called "high stakes" exams are not a valid measurement tool. Would you be able to send us some of the links to those—

Professor TOGNOLINI: What I said was that they decrease the validity. If you raise the stakes, it decreases the validity. It is not that they are not valid.

The CHAIR: Okay.

Professor TOGNOLINI: Because the HSC is high stakes.

The CHAIR: They are less valid. Can we get some references on that, if you could send those through to the secretariat?

Professor TOGNOLINI: I can give you a reference to that, yes.

The CHAIR: That would be helpful.

There were two references that I used in a presentation I gave in 2010. The presentation was entitled “Effective school leaders use information effectively to improve learning: An assessment perspective” and it was given at the Educational Leadership Conference at Wollongong on 26 February 2010.

The first reference was by Campbell in 1979:

The more any quantitative social indicator is used for social decision making, the more subject it will be to corruption pressure and more apt it will be to distort and corrupt the social pressures it is intended to monitor.

Campbell, Donald T (1979). "Assessing the impact of planned social change". Evaluation and Program Planning. ***2*** *(1): 67–90. doi:10.1016/0149-7189(79)90048-X*

The second reference was by George Madaus in 2002:

The higher the stakes involved in testing, the less likely you are to get an accurate measurement of the construct you most want to measure. So, you simply cannot have both high stakes and high validity because the higher the stakes the more corrupt the measure.

Unfortunately, I cannot locate the reference although I am happy to keep searching if required.

ANSWERS ON NOTICE 2: Page 13 of the transcript

Professor TOGNOLINI: One of the things I would like to finish off with is—I think you said at the start that the PISA data, et cetera, is saying that we are not succeeding as a system and we have got to change. If you actually look at the State's data—the official data at, say, year 12—you see that our performance has actually improved. This is where it is on curriculum, it is assessed by our State-based examinations et cetera, which are validated everywhere. It shows that in 2001 we had around about 5 per cent or 6 per cent of the kids working at the top level of Advanced English; now it is up around about 15 per cent, 13 per cent. We have got more students performing at higher levels now than we have ever had. They are the data that support it. Then you say that NAPLAN says—NAPLAN actually shows that we have got a flattening out.

The reason why we have a flattening out is not that the kids flatten out. It is that because the way NAPLAN was designed, we only have had a few items at that top level that we chose that we can improve. It is like trying to measure growth with a meter ruler, rather than millimetres, where we can show growth. But it does not say that we are getting worse. In terms of PISA, everybody says you cannot teach to PISA. So why are we judging our system on something we cannot teach to? There is a whole motivation factor associated with PISA. I do a lot of work in China. I do a lot of work in Hong Kong—Hong Kong is China—and Singapore et cetera.

The Hon. MATTHEW MASON-COX: Almost.

The Hon. ANTHONY D'ADAM: Not according to PISA.

Professor TOGNOLINI: Almost. I am probably a few years ahead of myself, but you know what I mean. I have to go there on Wednesday so I do not want to say anything wrong.

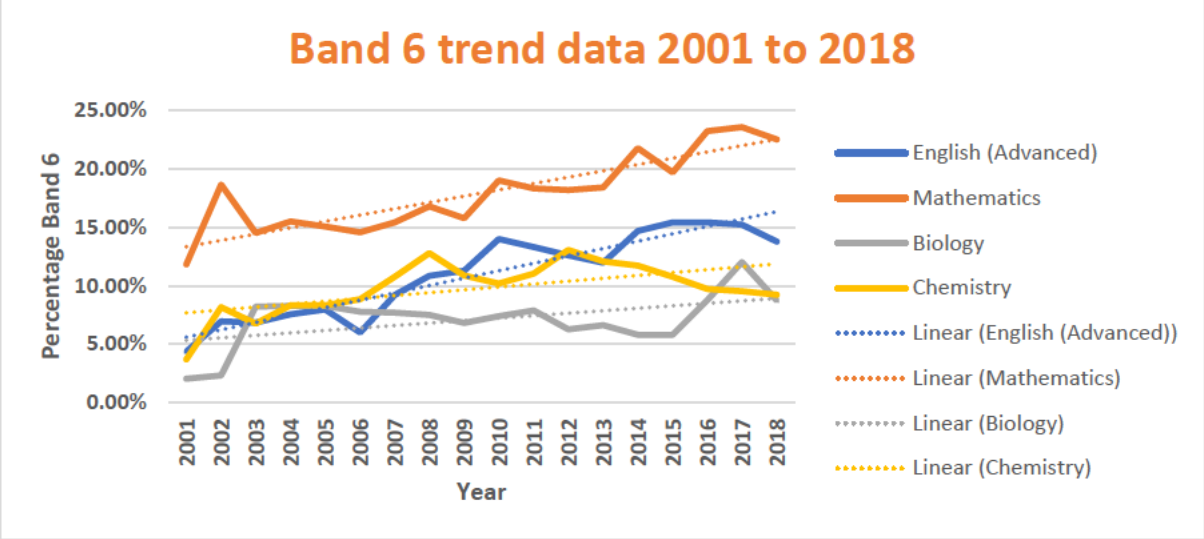
The CHAIR: No, you do not.

Professor TOGNOLINI: When they walk in, they walk in singing the national anthem. They are going to do it—represent their country. Our kids were, "Why are you picking on me to do this test? What do you mean it is not going—." That accounts for a huge number of marks. But we do not bother looking at it. Then if you say why is it going downwards within our own country, we can probably explain that too—I am sure we can. There is a whole demographic shift. The first year we did it we were motivated. What we have to do is look at the full range of data that are available before we start saying that our systems are failing. We want some other indicators. We do not notice because our systems are failing that people are not wanting to come to our universities. They think we are successful.

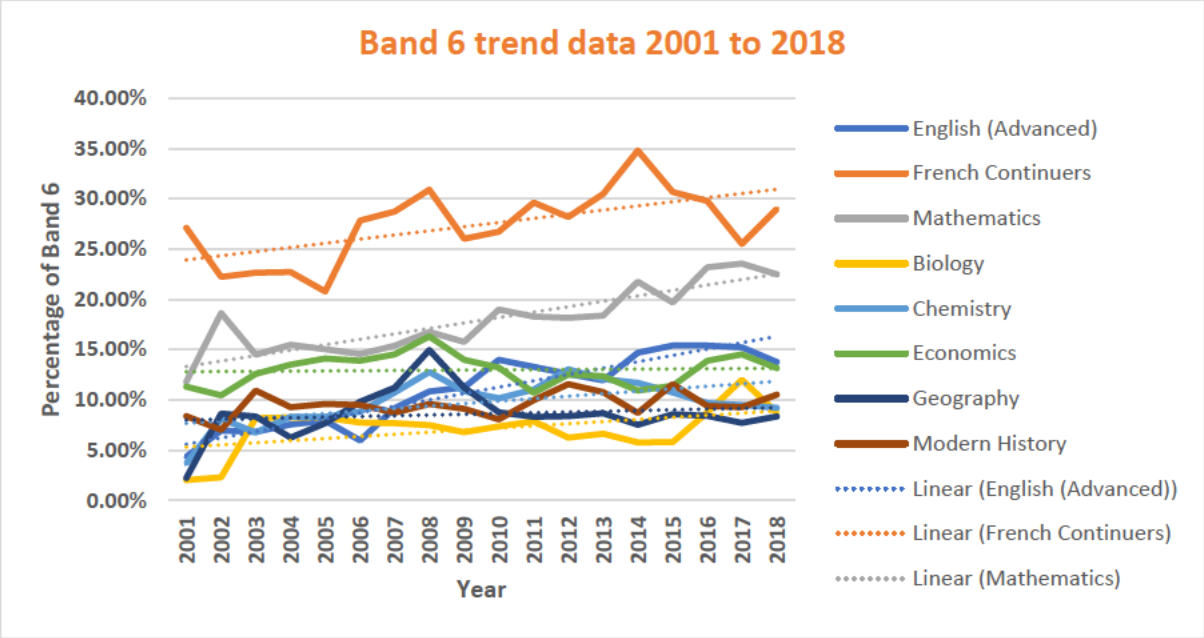
The CHAIR: On notice, can we get that data about year 12 because I am not too sure we have seen it as a Committee.

Professor TOGNOLINI: I can give you that.

The following graph shows the cross-temporal percentage of students in Band 6 for a small sample of subjects.



The following graph shows similar data for a broader range of subjects:



Attachment 1 provides

a. the proportions of students at Band 6 for a larger range of HSC subjects for the years 2001 to 2018;
b. the total number of students per subject, including all bands for the years 2001 to 2018; and,
c. the number of students in Band 6 for the same set of subjects for the years 2001 to 2018.

The data used to create the graphs and summary evidence on HSC Performance have been obtained from the NESA website and are available at:

https://www.boardofstudies.nsw.edu.au/bos_stats/hsc-pbds.html

Attachment 1

HSC Band 6

Proportion

Course Name	2001	2002	2003	2004	2005	2006	2007	2008	Calendar Year 2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
English (Advanced)	4.36%	6.95%	6.83%	7.56%	7.95%	5.99%	9.18%	10.83%	11.27%	13.98%	13.30%	12.58%	11.96%	14.68%	15.41%	15.41%	15.23%	13.77%
French Continuers	27.10%	22.25%	22.64%	22.72%	20.78%	27.83%	28.74%	30.90%	26.01%	26.72%	29.61%	28.18%	30.49%	34.79%	30.67%	29.78%	25.50%	28.97%
Mathematics	11.82%	18.63%	14.50%	15.50%	15.05%	14.56%	15.40%	16.77%	15.76%	19%	18.30%	18.17%	18.40%	21.76%	19.69%	23.20%	23.55%	22.50%
Biology	2.04%	2.33%	8.21%	8.25%	8.25%	7.77%	7.68%	7.49%	6.81%	7.38%	7.89%	6.26%	6.63%	5.79%	5.81%	8.76%	12.01%	8.74%
Chemistry	3.70%	8.14%	6.76%	8.33%	8.33%	8.84%	10.76%	12.79%	10.86%	10.17%	11.04%	13.05%	12.08%	11.71%	10.76%	9.70%	9.52%	9.22%
Economics	11.33%	10.46%	12.60%	13.53%	14.13%	13.89%	14.55%	16.32%	13.98%	13.22%	10.73%	12.52%	12.37%	10.93%	11.41%	13.91%	14.54%	13.17%
Geography	2.25%	8.66%	8.37%	6.26%	7.68%	9.81%	11.23%	14.97%	11.24%	8.76%	8.32%	8.39%	8.68%	7.53%	8.58%	8.42%	7.73%	8.35%
Modern History	8.40%	7.03%	10.93%	9.28%	9.59%	9.52%	8.73%	9.61%	9.13%	8.07%	9.99%	11.57%	10.77%	8.72%	11.58%	9.40%	9.29%	10.54%
Arabic Beginners	14.28%	0%	0%	0%	0%	0%	0%	0%	0%	0%	10%	0%	0%	0%	0%	0%	0%	16.66%
Arabic Continuers	25.95%	19.93%	15.16%	5.37%	2.62%	4.05%	7.29%	5.62%	2.84%	3.58%	5.17%	5.74%	8.08%	7.58%	10.43%	11.26%	9.81%	7.88%
Chinese Continuers	17.85%	48.00%	48.57%	51.06%	28.97%	32.67%	42.30%	35.29%	41.22%	43.22%	46%	50%	53.03%	27.71%	53.92%	44.89%	45.31%	45.08%
English (Standard)	0%	0%	0%	0%	0.01%	0.02%	0.07%	0.23%	0.20%	0.18%	0.16%	0.51%	0.41%	0.28%	0.37%	0.85%	0.85%	0.86%
French Beginners	13.34%	12.57%	13.11%	16.66%	17.16%	15.33%	18.86%	18.78%	16.63%	18.76%	17.65%	17.88%	16.33%	19.05%	21.63%	21.75%	22.30%	21.61%
Japanese Beginners	15.03%	18.04%	21.38%	19.96%	17.75%	23.12%	16.99%	15.50%	15.09%	17.33%	17.97%	16.26%	16.03%	13.39%	13.08%	17.59%	16.57%	14.20%
Japanese Continuers	16.99%	29.05%	28.36%	29.33%	28.10%	23.67%	25.63%	21.18%	22.47%	20.87%	24.43%	19.94%	13.54%	17.14%	18.96%	23.28%	28.12%	28.71%
Vietnamese Continuers	8.73%	9.02%	8.33%	0.96%	2.88%	2.56%	2.38%	2.02%	1.85%	2.17%	1.63%	0.64%	3.59%	5.71%	4.82%	3.20%	11.72%	3.54%

Total number of students per subject, including all bands

Course Name	2001	2002	2003	2004	2005	2006	2007	2008	Calendar Year 2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
English (Advanced)	20,145	20,893	24,603	27,004	27,581	27,767	28,115	27,459	27,281	27,163	27,138	27,244	27,039	26,730	26,006	26,080	26,780	26,129
French Continuers	760	755	808	836	871	873	842	851	888	883	834	887	823	799	727	779	698	635
Mathematics	20,891	20,213	19,939	19,830	19,125	18,219	17,826	17,308	17,271	17,216	16,626	16,740	16,536	16,694	16,451	16,139	17,060	17,826
Biology	12,455	12,284	12,301	13,026	13,269	14,140	14,495	15,311	15,342	15,915	16,773	16,628	16,950	17,138	17,271	17,735	18,153	18,106
Chemistry	9,017	8,925	9,380	10,187	10,179	10,256	10,335	10,193	10,092	10,387	11,026	10,883	11,084	11,173	10,907	10,554	10,974	11,134
Economics	5,496	5,179	5,204	5,874	5,567	5,456	5,716	5,432	6,163	6,141	5,431	5,262	5,302	5,131	5,090	5,196	5,261	5,191
Geography	6,254	6,127	5,577	4,423	4,930	4,524	4,539	4,314	4,572	4,621	4,431	4,325	4,133	4,418	4,276	4,283	4,589	4,427
Modern History	8,805	8,947	9,384	9,446	9,917	9,587	9,681	9,686	9,701	10,093	10,190	10,537	10,507	10,307	11,053	10,785	11,140	11,090
Arabic Beginners	7	7	5	6	0	5	0	1	0	9	10	5	3	0	1	3	1	6
Arabic Continuers	366	331	277	279	229	222	233	249	211	223	232	209	198	211	182	213	265	241
Chinese Continuers	56	75	70	94	107	101	130	85	131	118	100	62	66	83	102	98	128	173
English (Standard)	36,479	37,478	33,235	31,019	30,294	30,634	31,161	32,334	32,581	34,558	34,593	31,987	31,692	31,484	31,502	31,291	30,914	30,567
French Beginners	577	525	488	498	466	613	546	623	529	666	623	699	655	677	647	616	538	472
Japanese Beginners	326	327	449	581	552	588	606	774	762	669	534	621	630	687	642	665	712	718
Japanese Continuers	918	850	846	818	804	790	671	708	801	781	798	692	679	624	659	640	679	679
Vietnamese Continuers	126	144	132	104	104	117	126	148	162	184	183	155	139	140	145	125	145	141

No of students in Band 6

Course Name	2001	2002	2003	2004	2005	2006	2007	2008	Calendar Year 2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
English (Advanced)	878	1,452	1,680	2,042	2,193	1,663	2,581	2,974	3,075	3,797	3,609	3,427	3,234	3,924	4,008	4,019	4,079	3,598
French Continuers	206	168	183	190	181	243	242	263	231	236	247	250	251	278	223	232	178	184
Mathematics	2,469	3,766	2,891	3,074	2,878	2,653	2,745	2,903	2,722	3,271	3,043	3,042	3,043	3,633	3,239	3,744	4,018	4,011
Biology	254	286	1,010	1,075	1,095	1,099	1,113	1,147	1,045	1,175	1,323	1,041	1,124	992	1,003	1,554	2,180	1,583
Chemistry	334	727	634	849	848	907	1,112	1,304	1,096	1,056	1,217	1,420	1,339	1,308	1,174	1,024	1,045	1,027
Economics	623	542	656	795	787	758	832	887	862	812	583	659	656	561	581	723	765	684
Geography	141	531	467	277	379	444	510	646	514	405	369	363	359	333	367	361	355	370
Modern History	740	629	1,026	877	951	913	845	931	886	815	1,018	1,219	1,132	899	1,280	1,014	1,035	1,169
Arabic Beginners	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1
Arabic Continuers	95	66	42	15	6	9	17	14	6	8	12	12	16	16	19	24	26	19
Chinese Continuers	10	36	34	48	31	33	55	30	54	51	46	31	35	23	55	44	58	78
English (Standard)	0	0	0	0	3	6	22	74	65	62	55	163	130	88	117	266	263	263
French Beginners	77	66	64	83	80	94	103	117	88	125	110	125	107	129	140	134	120	102
Japanese Beginners	49	59	96	116	98	136	103	120	115	116	96	101	101	92	84	117	118	102
Japanese Continuers	156	247	240	240	226	187	172	150	180	163	195	138	92	107	125	149	191	195
Vietnamese Continuers	11	13	11	1	3	3	3	3	3	4	3	1	5	8	7	4	17	5

ANSWERS ON NOTICE 3: Page 14 of the transcript

The Hon. MATTHEW MASON-COX: If you could also provide the information on, if you like, the evidence that you are collecting to assess the capability of teachers into the system, so to speak.

Professor TOGNOLINI: Happy to, I guess. It is on the University of Sydney website.

The Hon. MATTHEW MASON-COX: You said it was being developed.

Professor TOGNOLINI: We are developing it now. This is for the things like creativity, cultural competency—we are doing all that now.

The Hon. SCOTT FARLOW: This is your assessment—

Professor TOGNOLINI: We are very happy to send what we have got.

There are 3 parts in response to this question.

1. Attachment 2 is a PDF of a Power Point presentation given in May 2019 which outlines the background to and the method by which the University of Sydney is intending to measure student performance on the 9 graduate outcomes identified in the University's strategic plan https://sydney.edu.au/dam/intranet/documents/strategy-and-planning/strategic-plan-2016-20.pdf

2. Attachment 3 is a PDF that contains the definitions and measurement rubrics for each of the graduate qualities. The validation process for these rubrics (measurement scales) is well underway and they have now been accepted in draft form by the Academic Board.

3. We are currently in the process of writing and publishing a set of academic papers to capture the link between policy and measurement; and, the psychometric theory that underpins the development of the measurement rubrics (scales) and the actual measurement of individual students on these scales.

I am happy to explain this process further if the Committee would like me to.

Attachment 2

An update on how we are going about measuring student performance on the University of Sydney's graduate qualities

Presented by

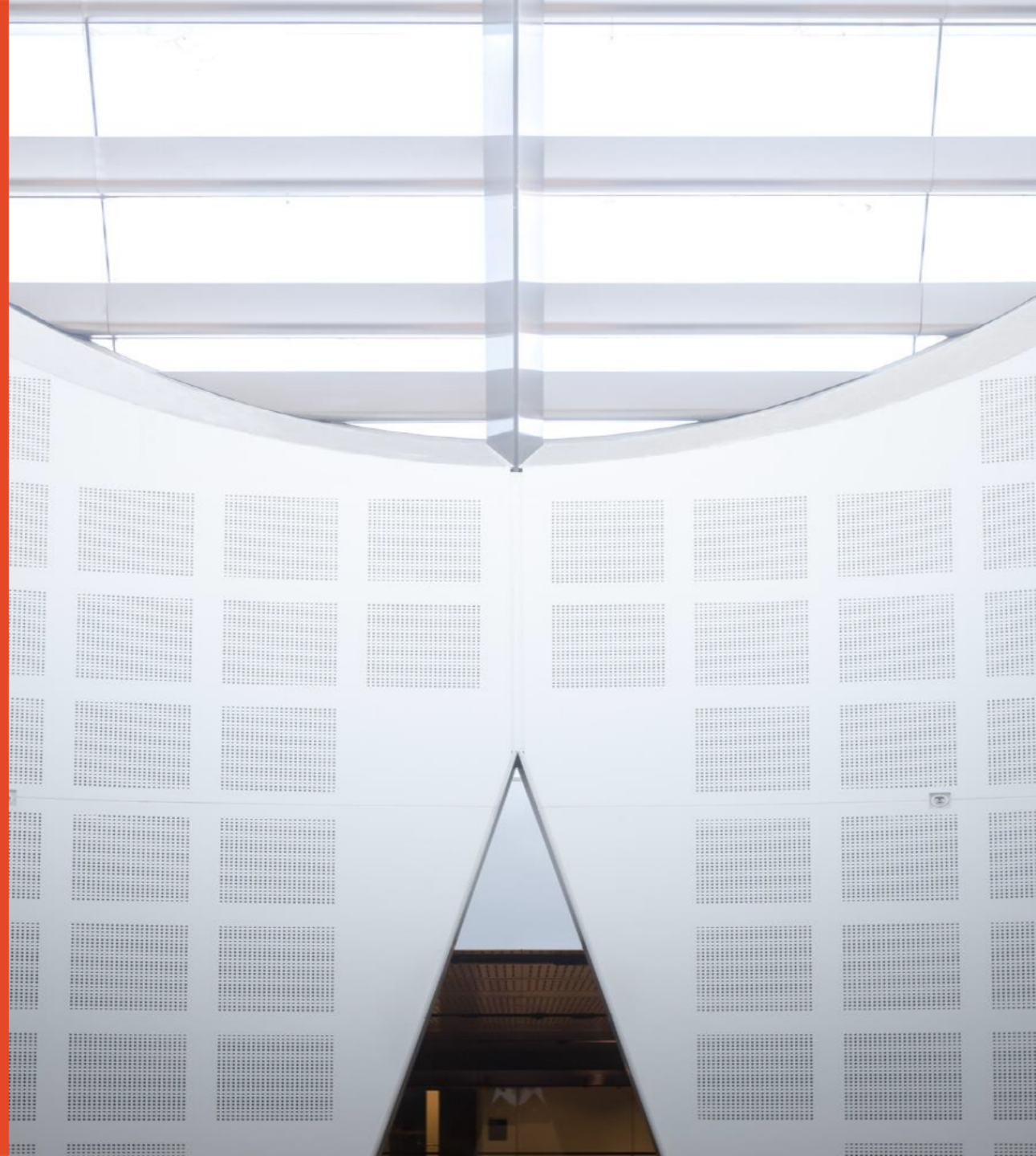
Jim Tognolini

Director of the Educational Measurement and Assessment Hub

University of Sydney, Australia

May 2019

THE UNIVERSITY OF SYDNEY



The context

University of Sydney Strategic Plan 2016-20

Strategy 4: Transform the undergraduate curriculum

4.1 Embed new graduate qualities and curriculum framework in all undergraduate degrees

- ***increase authentic and integrative assessment in each course component (minor, major, program and stream)***

University of Sydney Strategic Plan 2016-20

Strategy 4: Transform the undergraduate curriculum

4.1 Embed new graduate qualities and curriculum framework in all undergraduate degrees

- *increase authentic and integrative assessment in each course component (minor, major, program and stream)*

4.4 Develop a university-wide approach to assessing graduate qualities

- ***measure the attainment of graduate qualities from 2020***

University of Sydney Strategic Plan 2016-20

Strategy 4: Transform the undergraduate curriculum

4.1 Embed new graduate qualities and curriculum framework in all undergraduate degrees

- *increase authentic and integrative assessment in each course component (minor, major, program and stream)*

4.4 Develop a university-wide approach to assessing graduate qualities

- *measure the attainment of graduate qualities from 2020*

Delivering graduates with qualities that support first, second and third careers

University of Sydney Strategic Plan 2016-20

Strategy 5: Transform the learning experience

5.1 Develop interactive and collaborative learning designs that foster excellence and innovation

- ***design experiences that promote the alignment of learning and assessment at multiple levels (task, unit, major, degree) and across disciplines***
- ***consider mechanisms for assessment across multiple units, between disciplines and in interdisciplinary projects***

University of Sydney Strategic Plan 2016-20

Strategy 5: Transform the learning experience

5.1 Develop interactive and collaborative learning designs that foster excellence and innovation

- *design experiences that promote the alignment of learning and assessment at multiple levels (task, unit, major, degree) and across disciplines*
- *consider mechanisms for assessment across multiple units, between disciplines and in interdisciplinary projects*

Flexible, personalised and collaborative learning and assessment

University of Sydney Strategic Plan 2016-20

Strategy 5: Transform the learning experience

5.2 Create contemporary environments that enable flexible and interactive learning

- *reduce the volume of summative assessment and improve feedback to students and staff through increased low-stakes formative assessment*
- *assure the integrity of assessment as an integral component of the graduate qualities*

University of Sydney Strategic Plan 2016-20

Strategy 5: Transform the learning experience

5.2 Create contemporary environments that enable flexible and interactive learning

- *reduce the volume of summative assessment and improve feedback to students and staff through increased low-stakes formative assessment*
- *assure the integrity of assessment as an integral component of the graduate qualities*

Valuing authentic learning through assessments which develop reflective and autonomous learners

University of Sydney Strategic Plan 2016-20

Graduate qualities	
1.	Depth of disciplinary expertise
2.	Critical thinking and problem solving
3.	Communication (oral and written)
4.	Information and digital literacy
5.	Inventiveness
6.	Cultural competence
7.	Interdisciplinary effectiveness
8.	An integrated professional, ethical and personal identity
9.	Influence

University of Sydney Strategic Plan 2016-20

Broader skills

Graduate qualities
1. Depth of disciplinary expertise
2. Critical thinking and problem solving
3. Communication (oral and written)
4. Information and digital literacy
5. Inventiveness
6. Cultural competence
7. Interdisciplinary effectiveness
8. An integrated professional, ethical and personal identity
9. Influence

Strategic Plan implementation

Assessment Working Group has been established to:

Strategic Plan implementation

Assessment Working Group has been established to:

- develop common approach and techniques for assessing graduate qualities
- develop common approach to planning alignment and integration of assessment across course components (esp. majors)
- recommend common approach to assessment of collaborative, interdisciplinary and project-based learning
- recommend policy/course management options for integrating assessment across units of study, projects, etc.
- recommend policy reforms for reducing volume of summative assessment and making increased use of feedback to students and staff through formative assessment and learning analytics

Strategic Plan implementation

Assessment Working Group has been established to:

- **develop common approach and techniques for assessing graduate qualities**
- develop common approach to planning alignment and integration of assessment across course components (esp. majors)
- recommend common approach to assessment of collaborative, interdisciplinary and project-based learning
- recommend policy/course management options for integrating assessment across units of study, projects, etc.
- recommend policy reforms for reducing volume of summative assessment and making increased use of feedback to students and staff through formative assessment and learning analytics

Some of the challenges

Some of the challenges

1. **The University is intent on implementing the graduate qualities into its program so that students actually do improve (increase) the (amount) of quality that they have. "How do we measure HOW MUCH the University is impacting on the acquisition of student graduate qualities?" (That is how can we know that the graduates have MORE of each of these qualities when they graduate compared to when they enter the University?"**

Some of the challenges

1. The University is intent on implementing the graduate qualities into its program so that students actually do improve (increase) the (amount) of quality that they have. "How do we measure HOW MUCH the University is impacting on the acquisition of student graduate qualities?" (That is how can we know that the graduates have MORE of each of these qualities when they graduate compared to when they enter the University?"

2. **"How can we REPORT the level of achievement on each of these qualities on graduation on the official Record of Achievement for each and every student?"**

Some of the challenges

1. The University is intent on implementing the graduate qualities into its program so that students actually do improve (increase) the (amount) of quality that they have. "How do we measure the extent to which the University is impacting on the acquisition of student graduate qualities?" (That is how can we know that the graduates have MORE of each of these qualities when they graduate compared to when they enter the University?"

2. "How can we REPORT the level of achievement on each of these qualities on graduation on the official Record of Achievement for each and every student?"

Both of these questions are MEASUREMENT questions

My response

"If a thing exists, it exists in some amount. If it exists in some amount, it can be measured"

(Cronbach (1990).

Some theory

Theoretical foundations of a common approach to measurement

1. Define the construct (graduate qualities)

2. Construct (analytic/holistic) rubric to describe growth (progress) in what you want to measure

3. Build the evidential argument for validating the rubric as a legitimate measure

4. Construct assessment tasks to provide the evidence of what it is the students know, can do and "behave/feel/are" in relation to the rubric.

5. Measure the performance

Construct the measurement scale

1. Define the construct (graduate qualities)

2. Construct an analytic/holistic rubric to describe growth (progress) in what you want to measure

3. Build the evidential argument for validating the rubric as a legitimate measure

Measure the performance

4. Construct assessment tasks to provide the evidence of what it is the students know, can do and "behave/feel/are" in relation to the rubric.

5. Measure the performance

Constructing the measurement scale

Articulated university rubrics for assessing graduate qualities

University level rubrics for each of the graduate qualities

Articulated university rubrics for assessing graduate qualities

University level rubrics for each of the graduate qualities

Discipline specific rubrics for each of the graduate qualities

Discipline specific rubrics for each of the graduate qualities

Discipline specific rubrics for each of the graduate qualities

Articulated university rubrics for assessing graduate qualities

University level rubrics for each of the graduate qualities

Discipline specific rubrics for each of the graduate qualities

Discipline specific rubrics for each of the graduate qualities

Discipline specific rubrics for each of the graduate qualities

Assessment task rubric

Assessment task rubric

Assessment task rubric

Properties of the rubrics

University rubrics

1. describe performance expectations and proficiency levels in context of clear conceptual framework.

2. must be clear, detailed and complete; reasonable in scope; grounded in knowledge and affective domains.

3. must be elaborated so that curriculum, teaching and assessment are aligned.

4. facilitate development of learning outcomes, experiences and assessments that include graduate qualities.

Steps in constructing rubrics for measuring graduate qualities

Steps in building rubrics

Step 1: Define the construct/quality to be measured

Steps in building rubrics

Step 1: Define the construct/quality to be measured

Step 2: Decide on the components that represent the construct/quality

- **The components give the evidence for student performance specified in the standard and captured in the definition.**
- **The components must be as clear and unambiguous as possible.**
- **The number of components will depend on the construct/graduate quality being measured.**
- **The process of developing and refining components may be iterative and may involve numerous edits.**

Steps in building marking rubrics

Step 3: Develop descriptions of performance for each level of each component.

- **Describe the performance levels by using language that shows "growth" from low to high on each of the components (for an analytic rubric).**
- **Use descriptive language rather than evaluative judgements (e.g. excellent, very good, good, fair, poor). Evaluative judgements are not rubrics. They are old-fashioned grading scales.**
- **The performance levels must show increasing levels of performance quality.**

Examples of university level rubrics

Critical thinking and problem solving

Definition

Critical thinking and problem solving are the questioning of ideas, evidence and assumptions in order to propose and evaluate hypotheses or alternative arguments before formulating a conclusion or a solution to an identified problem.

Components

- Definition of problem or issue in context
- Critical questioning of ideas, evidence and assumptions
- Creation and evaluation of hypotheses or alternative arguments
- Formulation of defensible conclusions and best possible solutions.

Critical thinking and problem solving

Components	0	1	2	3	4
Definition of problem or issue in context		Describes the problem or issue.	Provides a basic definition of the problem or issue and shows that the problem or issue is situated in a context.	Provides an informative definition of the problem or issue, shows that the problem or issue is situated in a context, shows understanding of the main features of that context and explains why these matter, defines key terms, identifies desirable features of possible solutions.	Insightful and articulate. Analyses a context by consulting a suitably broad range of informational sources, identifies and appropriately frames a problem or issue within that context, gives a detailed and clear definition of the problem or issue, explains why this problem or issue matters, sets out criteria against which to measure possible solutions.

Critical thinking and problem solving

Components	0	1	2	3	4
Critical questioning of ideas, evidence and assumptions		Listens to and understands the ideas of others.	Recognises that ideas, evidence and assumptions need to be examined, shows awareness of differences in perspective, shows sensitivity to possible bias and error, seeks out those who have knowledge and expertise.	Questions received ideas, evidence and assumptions, engages with the work of genuine experts, critiques fallacious rhetoric, engages in rational argument, assesses currently available evidence, provides evidence to justify conclusions.	Open-minded and intellectually rigorous. Critically examines received ideas, evaluates the credibility and the methodology of experts, engages with competing views from various historical, intercultural and interdisciplinary perspectives, locates and assesses new evidence.

Critical thinking and problem solving

Components	0	1	2	3	4
Creation and evaluation of hypotheses or alternative arguments		Listens to and understands the ideas of others.	Recognises that ideas, evidence and assumptions need to be examined, shows awareness of differences in perspective, shows sensitivity to possible bias and error, seeks out those who have knowledge and expertise.	Questions received ideas, evidence and assumptions, engages with the work of genuine experts, critiques fallacious rhetoric, engages in rational argument, assesses currently available evidence, provides evidence to justify conclusions.	Open-minded and intellectually rigorous. Critically examines received ideas, evaluates the credibility and the methodology of experts, engages with competing views from various historical, intercultural and interdisciplinary perspectives, locates and assesses new evidence.

Critical thinking and problem solving

Components	0	1	2	3	4
Critical questioning of ideas, evidence and assumptions		Identifies and understands hypotheses put forward by others.	Recognises that current hypotheses and arguments may be suboptimal, assesses the existing hypotheses and arguments.	Generates new hypotheses and arguments, shows awareness of how they could be compared and tested, carries out these tests.	Creative and judicious. Generates original hypotheses and arguments, tests relevant hypotheses and arguments via reasoning, observation, or experiment, evaluates the results.

Critical thinking and problem solving

Components	0	1	2	3	4
Formulation of defensible conclusions and best possible solutions		Recognises conclusions and solutions offered by others.	Formulates basic solutions or conclusions.	Offers a solution or conclusion based on engagement with the relevant evidence, defends this solution or conclusion in light of relevant evaluative criteria.	Wise and decisive. Decides on the balance of the evidence, formulates conclusion or solution clearly in their own words, identifies the proper scope and significance of the conclusion commensurate with methods used, explains why this conclusion or solution is best when measured against relevant evaluative criteria.

Cultural Competence

Definition

Cultural Competence is the ability to actively, ethically, respectfully, and successfully engage across and between cultures. In the Australian context, this includes and celebrates Aboriginal and Torres Strait Islander cultures, knowledge systems, and a mature understanding of contemporary issues.

Components

- Awareness of one's own cultural values and worldview
- Actively seeking to understand norms and values of other cultures
- Ability to communicate across and between cultures

Cultural competence

Components	0	1	2	3	4
Awareness of one's own cultural values and worldview		Growing understanding of one's own cultural values, worldviews and practices: which may include emerging understanding of one's own culture through disciplinary or theoretical knowledge.	Recognises the importance of understanding one's own cultural norms and values	Supports cultural difference on a personal, group/institutional and society level.	Possesses deep and broad understanding of one's own, group, institutional and societal cultures, and promotes that understanding among others.

Cultural competence

Components	0	1	2	3	4
Understanding norms and values of other cultures: and ability to engage interculturally and cross culturally.		Seeks knowledge and understanding of the norms and values of different cultures, which may be through engagement with disciplinary knowledge or theory.	Identifies the advantages gained and barriers overcome through inter- and cross-cultural understanding and collaboration.	Adopts a position of critical cultural reflection, and investigates cultural change with humility and sensitivity, whether independently or through active listening or active sharing, as appropriate.	Applies extensive understanding of other cultures and the ability to collaborate within and across cultural boundaries to promote ethically just outcomes, as appropriate.

Cultural competence

Components	0	1	2	3	4
Ability to communicate across and between cultures		Recognises the need to listen and communicate sensitively in culturally diverse settings (i.e. listening, speaking, writing, presenting)	Demonstrates sensitive listening and communication in culturally diverse settings	Initiates thoughtful, accurate and respectful listening and communication with others in culturally diverse settings	Implements high-level communication skills and complex understandings of cultural differences through a range of techniques to interact with a variety of stakeholders

Validation of university level rubrics

Validation of university level rubrics

1. Focus groups (2019) – including review from experts across 5 countries
2. Stakeholder panels (2019)
3. Disciplinary evaluation (June 2019)
4. Assessment trials (2018-2019)

Validation exercise

1. In small groups on your table, consider some further activities that we might use to "validate" the university level rubrics.

2. Share your activities with others at the table.

Construct the measurement scale

1. Define the construct (graduate qualities)

2. Construct an analytic/holistic rubric to describe growth (progress) in what you want to measure

3. Build the evidential argument for validating the rubric as a legitimate measure

Measurement of performance against rubrics

Theoretical foundations of a common approach to measurement

1. Define the construct (graduate qualities)

2. Construct (analytic/holistic) rubric to describe growth (progress) in what you want to measure

3. Build the evidential argument for validating the rubric as a legitimate measure

4. Construct assessment tasks to provide the evidence of what it is the students know, can do and "behave/feel/are" in relation to the rubric.

5. Measure the performance

Measurement of performance

4. Construct assessment tasks to provide the evidence of what it is the students know, can do and “behave/feel/are” in relation to the rubric.

5. Measure the performance

Assessment of graduate qualities

1. We are NOT intending to have an omnibus assessment for all undergraduate students across the University e.g. no critical thinking test or cultural competence assessment given to all students

2. The intention is to collect evidence of student performance of each student in each graduate quality across a degree program (generally 4 years). Assessment plans will indicate which units are most appropriately designed to enable the assessment of each of the rubrics.

3. Lecturers in these units will be invited to construct “unit-specific assessment tasks” (including task rubrics) that will provide evidence that can be used to measure performance against the discipline specific rubric and the result referenced to the University-specific rubric.

Our challenge in assessing graduate qualities

1. Lecturers have to construct assessment tasks that will enable the students to provide "evidence" that can be used to **locate students on the rubrics** – this is a significant step from what happens at the moment, where lecturers generally write assessment tasks to assess whether students have attained the learning outcomes, but the rubrics generally describe the steps that the lecturer would carry out to arrive at the correct answer i.e. the rubric rewards students for providing the lecturer with the solution that the lecturer has in his/her mind.

2. As the results will be relatively high stakes, lecturers will eventually accountable for the quality of the assessment and the quality of the evidence that is used to locate the student along the measurement scale.

Cultural competence

Components	0	1	2	3	4
Awareness of one's own cultural values and worldview		Growing understanding of one's own cultural values, worldviews and practices: which may include emerging understanding of one's own culture through disciplinary or theoretical knowledge.	Recognises the importance of understanding one's own cultural norms and values	Supports cultural difference on a personal, group/institutional and society level.	Possesses deep and broad understanding of one's own, group, institutional and societal cultures, and promotes that understanding among others.

Our challenge in assessing graduate qualities

1. I keep asking the DVC “Do we really want to report performance at the level of students?”; She says to me “Can we do it?”; I usually reply “Theoretically we can?”; and, she says, “This is important, let’s keep going”.

Writing assessment tasks to fit the university/discipline level rubrics

1. In small groups on your table design a task (for whatever discipline you like) that will enable the students at the highest level of performance to demonstrate that they are at that level.

2. Share your activities with others at the table.

Where we are up to

Current rubric assessment trials

Unit of study coordinator	Area/unit of study	GQ rubric/s to be trialled
Martin Tomitsch (supported by James Meade)	ICPU Cambridge	• **Inventiveness** – assessed at a few points, 1 week in Sydney presentation, • **Interdisciplinary effectiveness** • **Communication** • **Cultural competence** • **Critical thinking and problem solving?**
Manjula Sharma	Physics	• **Communication** • **Disciplinary Depth** • **Critical Thinking**
Frances di Lauro	Professionalism in the Workplace OLE (OLEO2118)	• **An integrated personal/professional/ethical identity** • **Others, subject to discussion**

Current rubric assessment trials

Unit of study coordinator	Area/unit of study	GQ rubric/s to be trialled
Alice Williamson	Communication in Stem OLE (OLET1605)	• **Communication** • **Cultural competence**
Anthony Kadi	Professional Engineering Program in FEIT	• **?**
Matthew Pye	BIOL1006, AGEN3008 (SOLES)	• **Cultural Competence**
Inam Haq/Chris Roberts/Jane Conway	MD milestone projects 2-3 times per year FMH to pick a few clinical schools – pilot a different GQ rubric each Inam Haq to consult L&T and see if other volunteers	• **Interdisciplinary effectiveness (Health collaboration challenge in August)** • **TBA**

Thank you

Attachment 3

University Graduate Qualities and Common University Rubrics

Office of the Deputy Vice Chancellor (Education)

Legend

Performance indicators	**Level 0** No evidence available	**Level 1** Able to demonstrate application of given concepts, procedures and knowledge in straightforward contexts	**Level 2** In addition to level 1, able to demonstrate application of given concepts, procedures and knowledge in more complex contexts	**Level 3** In addition to level 1 and 2, able to demonstrate application of new concepts, procedures and knowledge in new and complex contexts	**Level 4** In addition to level 1, 2 and 3, able to demonstrate application, creation and integration of new concepts, procedures and knowledge at the highest level that could be envisaged.

The nine University Graduate Qualities

Depth of disciplinary expertise

Critical thinking and problem solving

Communication (oral and written)

Information and digital literacy

Inventiveness

Cultural competence

Interdisciplinary effectiveness

An integrated professional, ethical and personal identity

Influence

Depth of Disciplinary Expertise	
Definition	Deep disciplinary expertise is the ability to integrate and rigorously apply knowledge, understanding and skills of a recognised discipline defined by scholarly activity, as well as familiarity with evolving practice of the discipline.
Components	Understanding of conceptual space of recognised discipline Integration and rigorous application of disciplinary knowledge Awareness of the norms, culture and practice of the discipline Capabilities to participate in the evolving practice in the discipline

	0	1	2	3	4
Understanding of the content and boundaries of the discipline	Describes in general terms what the discipline involves.	Identifies broad foundational ideas and concepts using formal terminology and nomenclature associated with the discipline.	Outlines ideas and concepts from a range of different topics and associated skills within the discipline in some depth.	Describes the concepts, instruments and skills within the contemporary context of the discipline and map into a framework, at times appreciating areas of inconsistency.	Analyses the concepts and methodologies within the historical perspective and the contemporary context of the discipline and synthesises these into a coherent intellectual framework with appreciation of disciplinary gaps and limitations.
Application and integration of disciplinary knowledge	Demonstrates general awareness of the kinds of activities an individual operating in the discipline undertakes.	Formulates broad ideas about the appropriate application of disciplinary knowledge. Identifies evidence or data which is germane and relevant to activities which characterise their discipline.	Utilises knowledge and skills drawing on basic, discipline-specific tools in activities that characterise their discipline and explains their choice of strategies using an integrated approach.	Integrates knowledge and skills using discipline-specific tools in applying their knowledge to the activities that characterise their discipline, justifying their decisions. Connects disciplinary knowledge into an overarching internal disciplinary framework.	Weighs and integrates knowledge and skills using hands-on, instrumental or abstract tools in activities that characterise their discipline, including the justification and defence of their application of knowledge and skills. Connects disciplinary knowledge into an internal framework and is able to position that knowledge into the wider context within which their discipline sits.
Awareness of the norms, practices and culture of the discipline		Outlines in general terms the formal norms and informal practices which affect the way in which practitioners within a discipline operate.	Outlines the regulatory practices of the discipline demonstrating an understanding of the internal workings of its culture.	Exercises judgement within the regulatory practices of the discipline demonstrating understandings of the internal workings of the discipline; identifies actual and potential conflicts in the application and operation of cultural norms within the discipline.	Exercises nuanced judgement within the ethical and regulatory practices of the discipline demonstrating intricate understandings of the internal workings of the discipline in terms of the ways that it produces knowledge and artefacts, and how these are shared, assessed and accepted within the culture and practice of the discipline.
Capabilities to participate in the evolving practice in the discipline		Demonstrates awareness that disciplinary practice evolves, aware of broad historical changes which have occurred over time.	Analyses the ways in which disciplines evolve over time; supports analysis with relevant theoretical knowledge evidence and data.	Reviews knowledge that have led to differing perspectives and shares these while considering the interests and concerns of allied fields and disciplines.	Synthesises knowledge leading to expanded perspectives and insights, and negotiates the territories that the discipline shares with other fields. Advocates effectively to promote the evolution of disciplinary knowledge and practices in a range of contexts and situations.

Critical Thinking and Problem Solving	
Definition	Critical thinking and problem solving are the questioning of ideas, evidence and assumptions in order to propose and evaluate hypotheses or alternative arguments before formulating a conclusion or a solution to an identified problem.
Components	Definition of problem or issue in context Critical questioning of ideas, evidence and assumptions Creation and evaluation of hypotheses or alternative arguments Formulation of defensible conclusions and best possible solutions.

	0	1	2	3	4
Definition of problem or issue in context		Describes the problem or issue.	Provides a basic definition of the problem or issue, and shows that the problem or issue is situated in a context.	Provides an informative definition of the problem or issue, shows that the problem or issue is situated in a context, shows understanding of the main features of that context and explains why these matter, defines key terms, identifies desirable features of possible solutions.	Insightful and articulate. Analyses a context by consulting a suitably broad range of informational sources, identifies and appropriately frames a problem or issue within that context, gives a detailed and clear definition of the problem or issue, explains why this problem or issue matters, sets out criteria against which to measure possible solutions.
Critical questioning of ideas, evidence and assumptions		Listens to and understands the ideas of others.	Recognises that ideas, evidence and assumptions need to be examined, shows awareness of differences in perspective, shows sensitivity to possible bias and error, seeks out those who have knowledge and expertise.	Questions received ideas, evidence and assumptions, engages with the work of genuine experts, critiques fallacious rhetoric, engages in rational argument, assesses currently available evidence, provides evidence to justify conclusions.	Open-minded and intellectually rigorous. Critically examines received ideas, evaluates the credibility and the methodology of experts, engages with competing views from various historical, intercultural and interdisciplinary perspectives, locates and assesses new evidence.
Creation and evaluation of hypotheses or alternative arguments		Identifies and understand hypotheses put forward by others.	Recognises that current hypotheses and arguments may be suboptimal, assesses the existing hypotheses and arguments.	Generates new hypotheses and arguments, shows awareness of how they could be compared and tested, carries out these tests.	Creative and judicious. Generates original hypotheses and arguments, tests relevant hypotheses and arguments via reasoning, observation, or experiment, evaluates the results.
Formulation of defensible conclusions and best possible solutions		Recognises conclusions and solutions offered by others.	Formulates basic solutions or conclusions.	Offers a solution or conclusion based on engagement with the relevant evidence, defends this solution or conclusion in light of relevant evaluative criteria.	Wise and decisive. Decides on the balance of the evidence, formulates conclusion or solution clearly in their own words, identifies the proper scope and significance of the conclusion commensurate with methods used, explains why this conclusion or solution is best when measured against relevant evaluative criteria.

Communication (oral and written)	
Definition	Effective communication, in both oral and written form, is the clear exchange of meaning in a manner that is appropriate to audience and context.
Components	Clear conveyance of meanings in terms original to the student Adjustment according to audience and context Use of media and modes appropriate to each communication Clarity of structure and organization of ideas

	0	1	2	3	4
Communicates meaning in own words or 'voice'		Communicates meaning which for the most part clearly and accurately distinguishes own voice from that of external sources.	Accurately paraphrases and summarises meaning using own voice.	Communicates meaning unambiguously in their own voice, while integrating information from multiple sources to present alternative cases.	Communicates meaning skillfully and unambiguously in their own voice while synthesising and integrating information from multiple and conflicting sources
Adjusts communication according to context (situation, audience, purpose and genre)		Adjusts communication in a manner that demonstrates awareness of given context.	Adjusts communication in a manner that demonstrates awareness of different contexts.	Adjusts communication in a manner that demonstrates sensitivity to a given context	Adjusts communication in a nuanced manner, demonstrating sensitivity to given context demonstrated in communicative style
Uses different modes, media and technology according to context		Uses different modes, media and technology in communication appropriately.	Uses a variety of appropriate modes, media and technology in communication to promote understanding and engagement.	Distinguishes between different modes, media and technology to enhance communication and to promote understanding and engagement.	Distinguishes between and uses different and appropriate modes, media and technology inventively to enhance communication and to enrich understanding and engagement
Structures and organises ideas and information according to context		Structures and organises ideas and information logically	Structures and organises ideas, and information logically and clearly	Structures and organises ideas, and logically, clearly and cohesively	Structures and organises ideas persuasively, and information consistently with clarity, cohesion and logic

Information and Digital Literacy	
Definition	Information and digital literacy is the ability to locate, interpret, evaluate, manage, adapt, integrate, create and convey information using appropriate resources, tools and strategies.
Components	Location, interpretation and evaluation of data and information Management of data and information Adaptation, integration and conveyenace of data and information Creation of data and information Effective use of digital resources, tools and strategies

	0	1	2	3	4
Scope of an information need		Identifies main concepts when researching a straightforward question or problem, with minimal reference to context.	Uses the context of an information need to inform its scope	Adapts approaches from multiple disciplines and uses them in more complex/specialised contexts	Produces novel insights and approaches.
Location of data and information		Applies commonly used search tools and strategies provided to access and select data and information	Evaluates a variety of search strategies and sources and selects an appropriate set of these to use	Makes sophisticated use of search strategies and sources appropriate to a disciplinary context	Critiques and creates well-designed search strategies and makes innovative choices of sources
Interpretation and evaluation of sources		Applies basic criteria provided to judge the appropriateness of data and information and gives meaning within a defined context	Independently applies basic criteria to judge the value of information in a disciplinary context	Adapts criteria recognised within disciplines to judge the appropriateness of data and information and extracts multiple meanings.	Creates and justifies innovative criteria to judge the appropriateness of data and information and systematically constructs insightful meanings from multiple perspectives.
Adaptation, integration and synthesis		Uses basic techniques to extract and organise information and data	Selects and applies basic extraction and synthesis techniques to organise more complex information	Extracts information from multiple sources, and, organises and synthesises it coherently to satisfy a clear purpose	Extracts information in innovative ways, and, organises and synthesises data to create new knowledge.
Use of digital resources, tools, and strategies		Uses basic digital tools and strategies in simple ways under close supervision and guidance	Uses basic and intermediate digital tools and strategies in simple ways with minimal supervision and guidance	Applies best practice approaches when using digital tools and strategies and shows evidence of independently learning to use new and more sophisticated techniques	Evaluates and uses advanced features of digital tools in sophisticated ways and shows evidence of independently learning to use a diverse range of new tools and strategies in innovative ways.
Ethical and legal access and use of data and information		Follows ethical, legal and disciplinary standards under close guidance and supervision in sourcing data and information at a basic level to cite sources and indicate direct reuse	Independently follows ethical, legal and disciplinary standards in sourcing data or information at a basic level to cite sources and indicate direct reuse	Identifies and resolves ethical dilemmas in sourcing and interpreting data or information	Identifies ethical dilemmas in sourcing data or information and evaluates them using multiple frameworks in order to comply with ethical, legal and disciplinary standards.

Inventiveness	
Definition	Inventiveness is generating novel ideas and solutions.
Components	Reimagines and reframes disparate ideas, observations or resources Creates novel, ideas, solutions or actions.

	0	1	2	3	4
Creative thinking: coming up with ideas and using resources		Generates one-dimensional ideas and/or adopts resources within disciplinary norms and conventions.	Generates and connects similar ideas, and adopts resources within disciplinary norms and conventions.	Generates, connects and synthesises multiple ideas, and uses resources outside disciplinary norms and conventions.	Generates, connects and synthesises disparate ideas, and draws on resources in a way that demonstrates the ability to transcend and move between disciplinary norms and conventions.
Process and strategy: implementing a plan* * ***Might not apply to all disciplines***		Follows a strategy that is identical with previously documented processes, and/or executes a plan that follows pre-set steps.	Follows an organised strategy that uses a combination of previously documented processes, and/or executes a plan that allows for flexibility and adaptation.	Follows an organised strategy that draws on previously documented processes, and a reflective execution of a plan that allows for flexibility and adaptation.	Follows an organised strategy that goes beyond previously documented processes, and reflective execution and evaluation of a plan that allows for flexibility and adaptation.
Outputs: developing concepts, solutions, processes or actions		Creates outputs that are a copy to something existing, incomplete, not feasible and/or poorly contextualised.	Creates outputs that show original aspects, and/or that are mostly resolved, practical and/or contextualised.	Creates outputs that are original, and/or that are resolved, feasible and appropriately contextualised.	Creates outputs that are original, resolved, feasible and contextualised in unique and novel ways.

Cultural Competence	
Definition	Cultural Competence is the ability to actively, ethically, respectfully, and successfully engage across and between cultures. In the Australian context, this includes and celebrates Aboriginal and Torres Strait Islander cultures, knowledge systems, and a mature understanding of contemporary issues.
Components	Awareness of one's own cultural values and worldview Actively seeking to understand norms and values of other cultures

	0	1	2	3	4
Awareness of one's own cultural values and worldview		Growing understanding of one's own cultural values, worldviews and practices: which may include emerging understanding of one's own culture through disciplinary or theoretical knowledge.	Recognises the importance of understanding one's own cultural norms and values	Supports cultural difference on a personal, group/institutional and society level.	Possesses deep and broad understanding of one's own, group, institutional and societal cultures, and promotes that understanding among others.
Understanding norms and values of other cultures: and ability to engage interculturally and cross culturally.		Seeks knowledge and understanding of the norms and values of different cultures, which may be through engagement with disciplinary knowledge or theory.	Identifies the advantages gained and barriers overcome through inter- and cross-cultural understanding and collaboration.	Adopts a position of critical cultural reflection, and investigates cultural change with humility and sensitivity, whether independently or through active listening or active sharing, as appropriate.	Applies extensive understanding of other cultures and the ability to collaborate within and across cultural boundaries to promote ethically just outcomes, as appropriate.
Ability to communicate across and between cultures		Recognises the need to listen and communicate sensitively in culturally diverse settings (i.e.listening, speaking, writing, presenting)	Demonstrates sensitive listening and communication in culturally diverse settings	Initiates thoughtful, accurate and respectful listening and communication with others in culturally diverse settings	Implements high-level communication skills and complex understandings of cultural differences through a range of techniques to interact with a variety of stakeholders

Interdisciplinary effectiveness	
Definition	Interdisciplinary effectiveness is the integration and synthesis of multiple viewpoints and practices, working effectively across disciplinary boundaries.
Components	Understanding of multiple viewpoints and practices Working effectively across discipline and professional boundaries Integrating and synthesising different ways of thinking Production of distinctive outcomes.

	0	1	2	3	4
Understanding of multiple viewpoints and practices		Recognises and acknowledges different roles and viewpoints within an interdisciplinary team.	Considers likely boundaries, biases, ideas, criticisms and amendments contributed by other disciplines when addressing complex problems.	Articulates problem solving approaches by incorporating knowledge and perspectives within and across disciplines.	Enacts ones' discipline-based academic and/or professional responsibilities while appreciating the diversity of knowledge from the wider community and disicplines.
Integrating and synthesising different ways of thinking		Demonstrates receptivity, flexibility, and willingness to integrate new knowledge, skills, and behaviours as contributed by several disciplines.	Displays sensitivity, empathy, trust and commitment towards other's roles/ positions in collective problem-solving.	Critically analyses and displays insights on one's own as well as team's strengths and limitations when contributing to the team's collaborative practice to achieve solutions to complex outcomes.	Creatively adapts in their contribution to the team's collaborative practice in order to achieve shared solutions to complex outcomes.
Working effectively across discipline and professional boundaries		Respectfully conducts oneself when identifying potential sources of conflict when working with other disciplines	Seeks opinions, and provides timely, sensitive and constructive feedback to colleagues in the context of team culture.	Engages with a willingness to find a compromise between and within disciplines; including respectful conflict resolution where appropriate.	Displays situational leadership: Understands, interacts, manages and adjusts behaviour of self and others to achieve common goals.
Production of distinctive outcomes.		Contributes towards developing a shared goal, and in negotiating the achievement of unified plan and distinctive outcomes.	Actively applies principles of collaboration in negotiating goals, plans and outcomes.	Engages in planning a collaborative solution whilst accommodating team's strengths, limitations, and opportunities.	Evaluates critical success factors in proposing solutions to the defined complex problem.

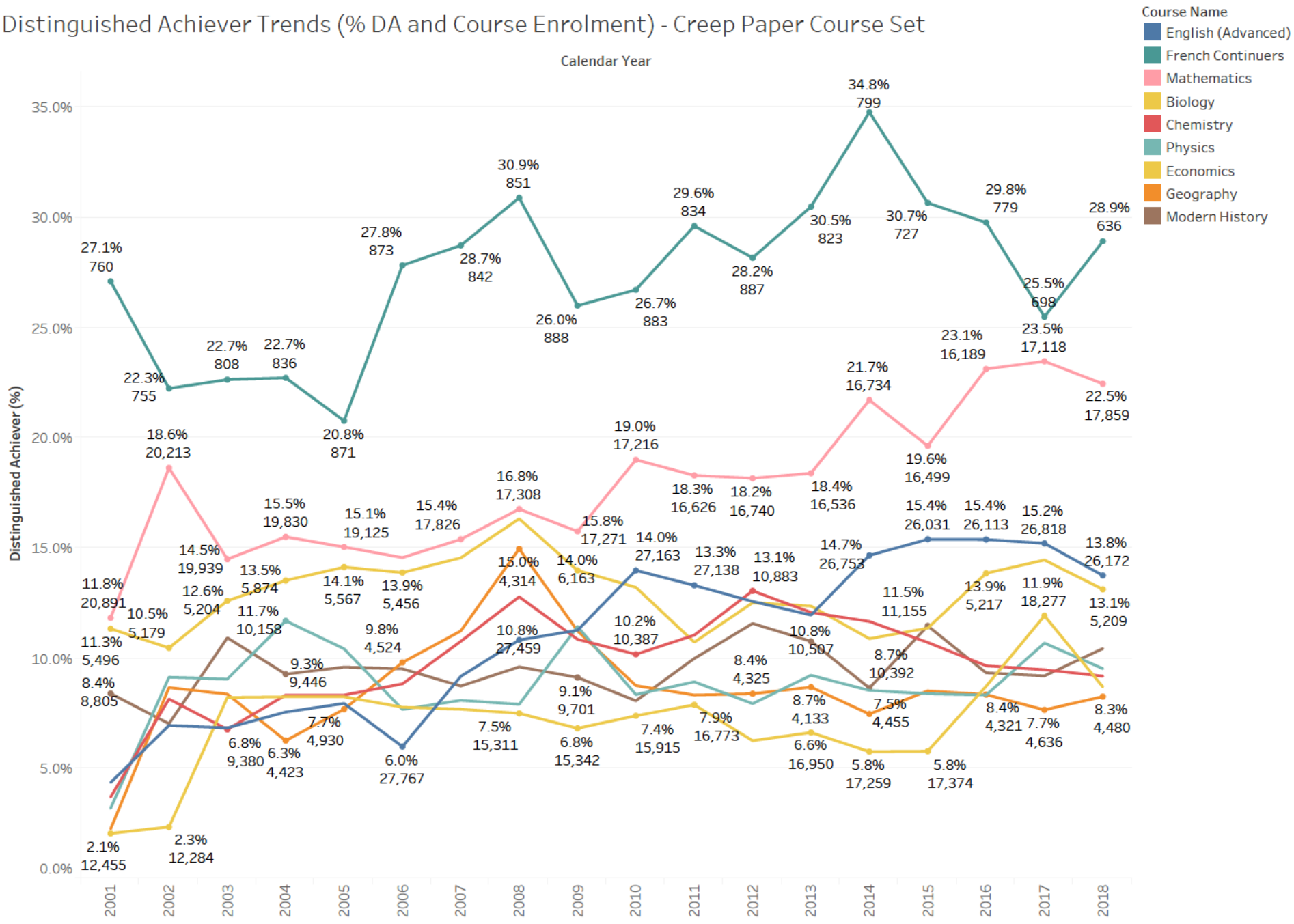
An integrated professional, ethical and personal identity	
Definition	An integrated professional, ethical and personal identity is understanding the interaction between one's personal and professional selves in an ethical context.
Components	Articulates a coherent ethical framework Reflects on the self in personal and professional contexts

	0	1	2	3	4
Articulation of ethical values and practices		Ability to identify core values of ethical conduct including, for example, justice, beneficence, integrity and respect for all human and non-human beings and the environment, and to describe where they may be relevant. Awareness of what it is to be ethical or not ethical and demonstrates capacity to contrast the ethical with the not ethical in specific contexts.	Ability to engage with core values of ethical conduct and identify the relevant issues that require consideration in a specific context/decision e.g. relevance of, and need for consent, confidentiality, disclosure, inter-cultural and intra-cultural agreement. Demonstrates ability to reflect on values, value-conflicts, and different views/positions that others may hold.	Demonstrates ability to think critically and can provide reasons for choices and actions with reference to core values of ethical conduct. Shows evidence that alternative views have been considered in own reasoning and decisions.	Ability to identify, articulate and respond with regard to all the relevant ethical considerations in any given context – providing clear reasons for decisions and actions. Demonstrates appreciation of different perspectives, and roles, and the need to consider the value of alternative views/perspectives and how understanding the views of others allows us to develop and formulate our own ethical identity.
Responsibilities		Awareness of the need to take responsibility for actions. Can give examples of specific actions that might/should/would be taken.	Takes responsibility for decisions and actions.	Takes responsibility for decisions and actions – taking into account the impact on other individuals.	Takes responsibility for decisions and actions – taking into account the impact on other individuals, society and the environment.
Articulation of ethical values and practices in professional contexts		Awareness of role-specific/professional ethical responsibilities	Awareness of role-specific/professional ethical responsibilities and is aware of the sources of these.	Awareness of role-specific/professional ethical responsibilities and demonstrates capacity to describe the source/s of these.	Ability to articulate role-specific/professional ethical responsibilities and demonstrates capacity to critique the source/s of these.

Influence	
Definition	Influence is engaging others in a process, idea or vision.
Components	Responsibility for improvement through involvement and leadership Confidence, self-awareness and a willingness to learn from others Persuasiveness

	0	1	2	3	4
Confidence and self-efficacy in leading others		Understands themselves and their own abilities. Expresses own opinions when prompted.	Expresses own opinions without prompting. Shows capacity to understand others and how their actions may impact them.	Confidently attempts to influence others with an understanding of how their actions may impact others. Responds to new challenges. Able to reflect on their own leadership.	Leads with confidence and seeks out opportunities to lead others Initiates reflection on leadership skills and puts in place strategies for self-development and successfully responding to challenges.
Willingness to engage with, learn from and understand others		Engages with others. Listens to others.	Will initiate tasks, engage with or learn from others in their own discipline.	Completes tasks and engages with and guides others within their discipline when directed. Attempts to identify the skills and needs of others and recognise their potential to contribute to shared learning. Considers a range of viewpoints.	Initiates and accepts accountability for tasks. Understands clearly what distinct knowledge may be learned from others and negotiates with others to take on relevant tasks. Mentors or empowers others to reach their potential. Actively seeks out opportunities to engage with others on a range of issues both within and external to their expertise. Seeks out new and diverse viewpoints and resources.
Contextually relevant persuasion.		Understands ethical persuasion.	Interprets the social context in which persuasion is required.	Persuades ethically, with knowledge of the social context, the beliefs, attitudes, motivations and/or behaviours of others.	Persuades with a clear understanding of their own ethical perspective, the relevant ethical framework for the situation and the perspectives of others. Reflects on the impact that persuasive actions have on those around them and the wider society.
Effective techniques of persuasion.		Uses their own opinion in attempting to persuade. Uses structured arguments for persuasion.	When persuading, uses opinions of from themselves and others without providing reference or context. Can identify an appropriate audience. Arguments exhibit logic.	Persuasion supported by reference to evidence and/or the opinions of experts. Understands their audience and can identify an appropriate communication channel. Persuades with arguments that are coherent and have logical flow.	Persuades using high quality evidence including the opinions of experts and people with lived experience. Persuades using, where relevant, a range of appropriate communication channels. Persuades using arguments that are coherent, flow logically and synthesise relevant evidence.

Attachment 4



Distinguished Achiever Trends (% DA)

Difference in Band 6 (%)
-5.00 15.00

	Calendar Year																	
Course Name	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
English (Advanced)	4.4%	7.0%	6.8%	7.6%	8.0%	6.0%	9.2%	10.8%	11.3%	14.0%	13.3%	12.6%	12.0%	14.7%	15.4%	15.4%	15.2%	13.8%
French Continuers	27.1%	22.3%	22.7%	22.7%	20.8%	27.8%	28.7%	30.9%	26.0%	26.7%	29.6%	28.2%	30.5%	34.8%	30.7%	29.8%	25.5%	28.9%
Mathematics	11.8%	18.6%	14.5%	15.5%	15.1%	14.6%	15.4%	16.8%	15.8%	19.0%	18.3%	18.2%	18.4%	21.7%	19.6%	23.1%	23.5%	22.5%
Biology	2.1%	2.3%	8.2%	8.3%	8.3%	7.8%	7.7%	7.5%	6.8%	7.4%	7.9%	6.3%	6.6%	5.8%	5.8%	8.7%	11.9%	8.7%
Chemistry	3.7%	8.2%	6.8%	8.3%	8.3%	8.8%	10.8%	12.8%	10.9%	10.2%	11.1%	13.1%	12.1%	11.7%	10.7%	9.7%	9.5%	9.2%
Physics	3.2%	9.1%	9.1%	11.7%	10.4%	7.7%	8.1%	7.9%	11.4%	8.4%	8.9%	7.9%	9.2%	8.5%	8.4%	8.3%	10.7%	9.5%
Economics	11.3%	10.5%	12.6%	13.5%	14.1%	13.9%	14.6%	16.3%	14.0%	13.2%	10.7%	12.5%	12.4%	10.9%	11.4%	13.9%	14.5%	13.1%
Geography	2.3%	8.7%	8.4%	6.3%	7.7%	9.8%	11.2%	15.0%	11.2%	8.8%	8.3%	8.4%	8.7%	7.5%	8.5%	8.4%	7.7%	8.3%
Modern History	8.4%	7.0%	10.9%	9.3%	9.6%	9.5%	8.7%	9.6%	9.1%	8.1%	10.0%	11.6%	10.8%	8.7%	11.5%	9.3%	9.2%	10.4%

Distinguished Achiever Trends - Delta 2003

2003 was chosen as reference year given calibration changes in 2001-2002

	Calendar Year															
Course Name	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
English (Advanced)	0.0%	0.7%	1.1%	-0.9%	2.3%	4.0%	4.4%	7.2%	6.5%	5.7%	5.1%	7.8%	8.6%	8.6%	8.4%	6.9%
French Continuers	0.0%	0.1%	-1.9%	5.2%	6.1%	8.3%	3.4%	4.1%	7.0%	5.5%	7.9%	12.1%	8.0%	7.1%	2.9%	6.3%
Mathematics	0.0%	1.0%	0.6%	0.1%	0.9%	2.3%	1.3%	4.5%	3.8%	3.7%	3.9%	7.2%	5.1%	8.6%	9.0%	8.0%
Biology	0.0%	0.0%	0.0%	-0.4%	-0.5%	-0.7%	-1.4%	-0.8%	-0.3%	-2.0%	-1.6%	-2.5%	-2.4%	0.5%	3.7%	0.5%
Chemistry	0.0%	1.6%	1.6%	2.1%	4.0%	6.0%	4.1%	3.4%	4.3%	6.3%	5.3%	4.9%	4.0%	2.9%	2.7%	2.4%
Physics	0.0%	2.6%	1.4%	-1.4%	-1.0%	-1.2%	2.4%	-0.7%	-0.1%	-1.1%	0.2%	-0.5%	-0.7%	-0.7%	1.6%	0.5%
Economics	0.0%	0.9%	1.5%	1.3%	2.0%	3.7%	1.4%	0.6%	-1.9%	-0.1%	-0.2%	-1.7%	-1.2%	1.3%	1.9%	0.5%
Geography	0.0%	-2.1%	-0.7%	1.4%	2.9%	6.6%	2.9%	0.4%	0.0%	0.0%	0.3%	-0.9%	0.2%	0.0%	-0.7%	-0.1%
Modern History	0.0%	-1.7%	-1.3%	-1.4%	-2.2%	-1.3%	-1.8%	-2.9%	-0.9%	0.7%	-0.2%	-2.3%	0.5%	-1.6%	-1.7%	-0.5%

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. Proceedings of the 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.

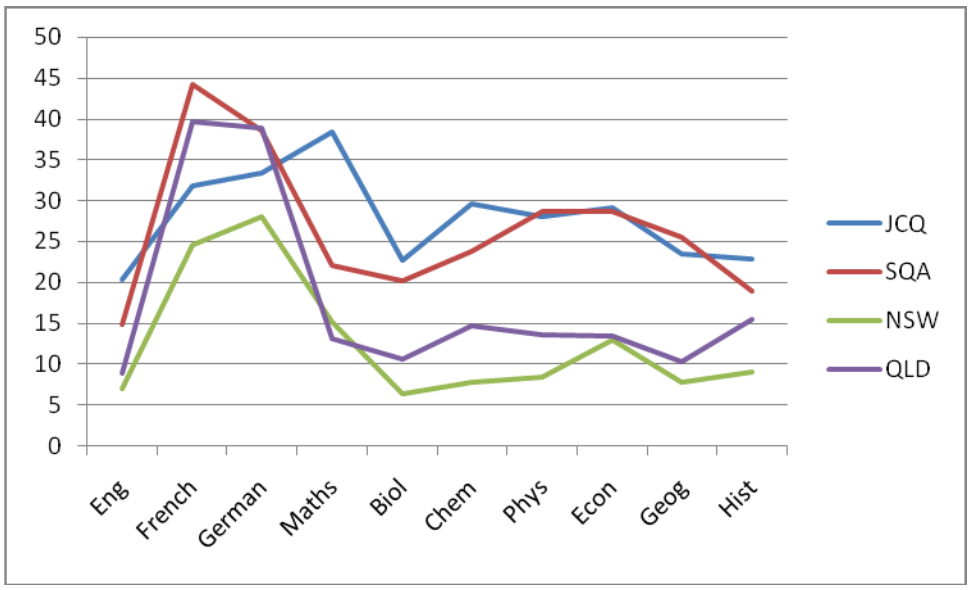


Figure 1: Average percentage of students in top grade for each authority across 10 subjects

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.

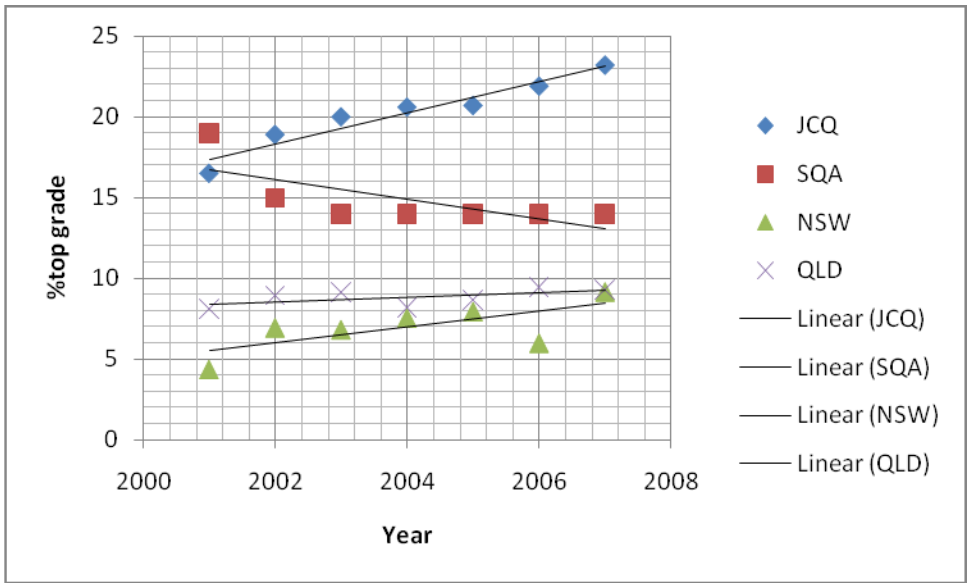


Figure 2: Trend for English top percentage

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.

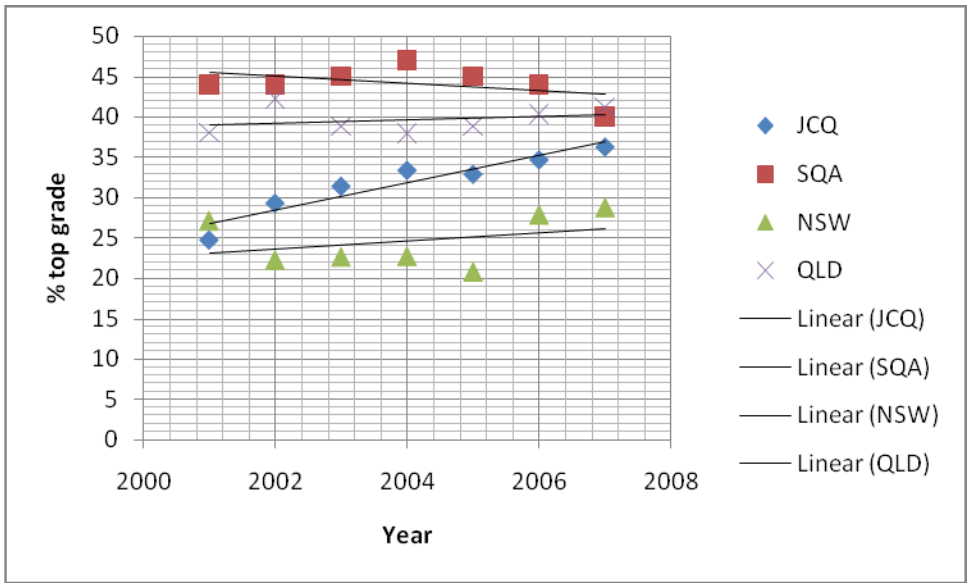


Figure 3: Trend for French top percentage

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.

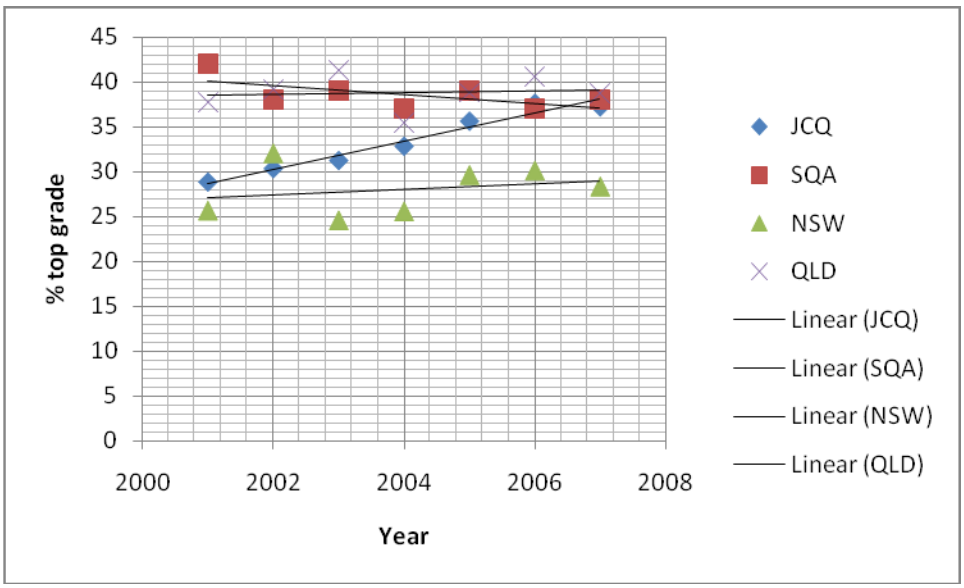


Figure 4: Trend for German top percentage

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

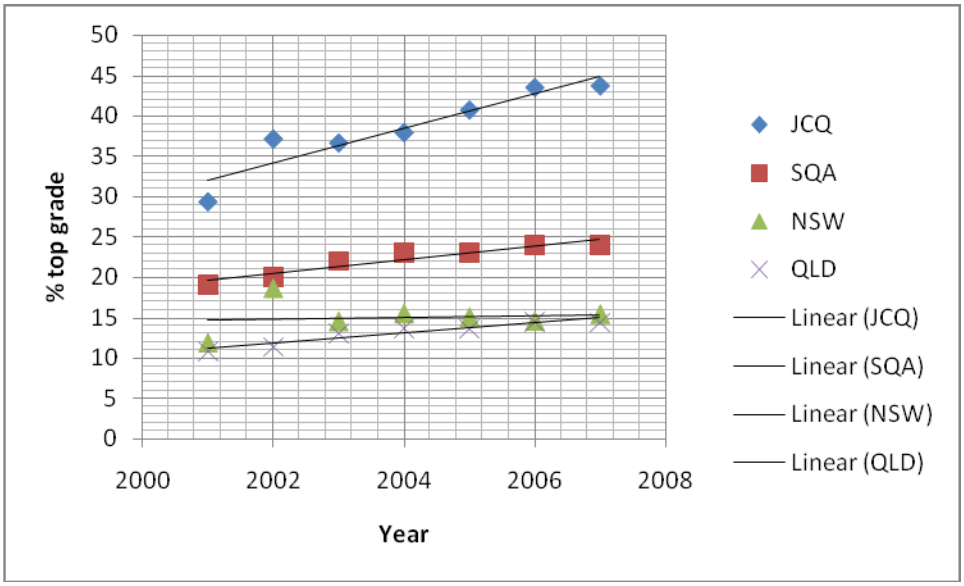


Figure 5: Trend for Maths top percentage

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

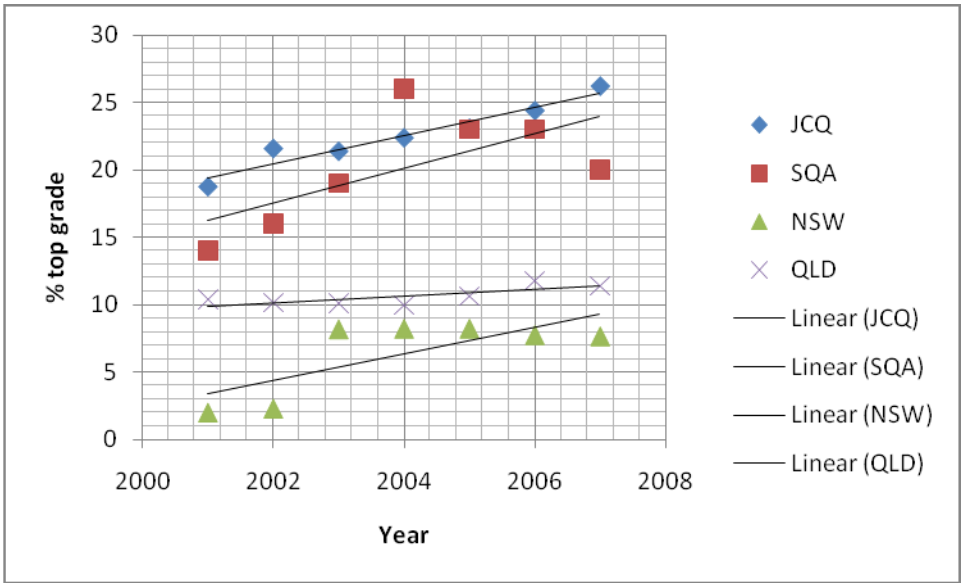


Figure 6: Top percentage for Biology

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

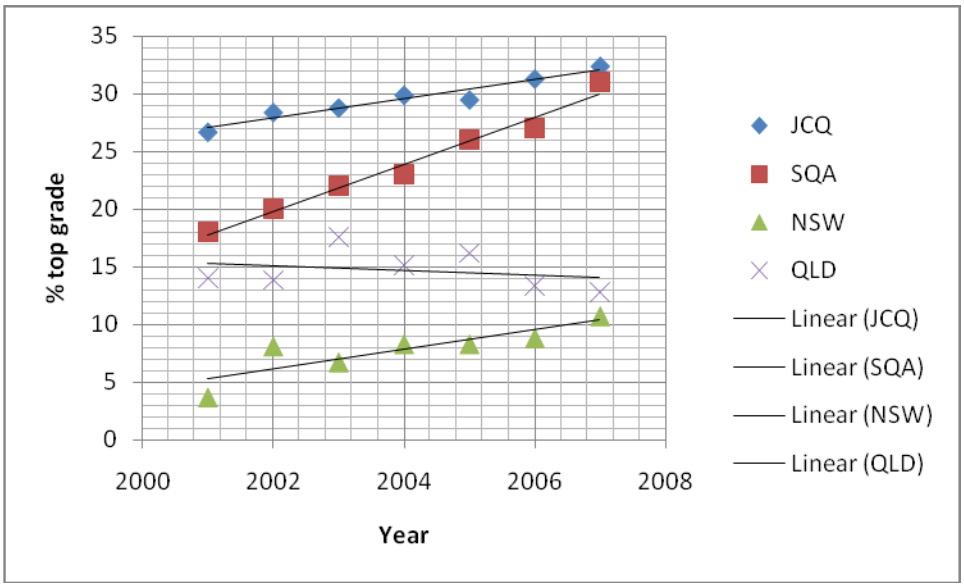


Figure 7: Top percentage for Chemistry

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

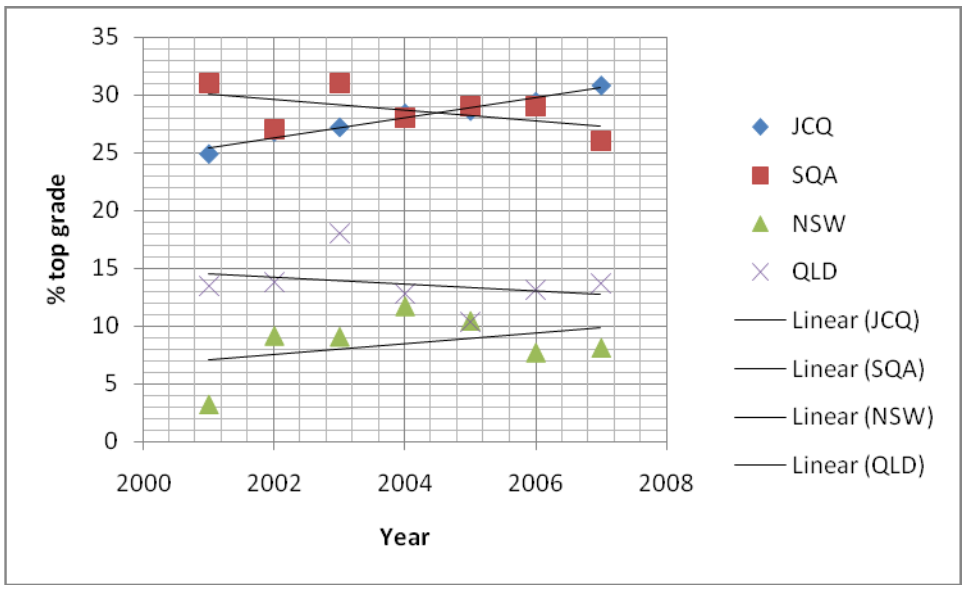


Figure 8: Top percentage for Physics

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.

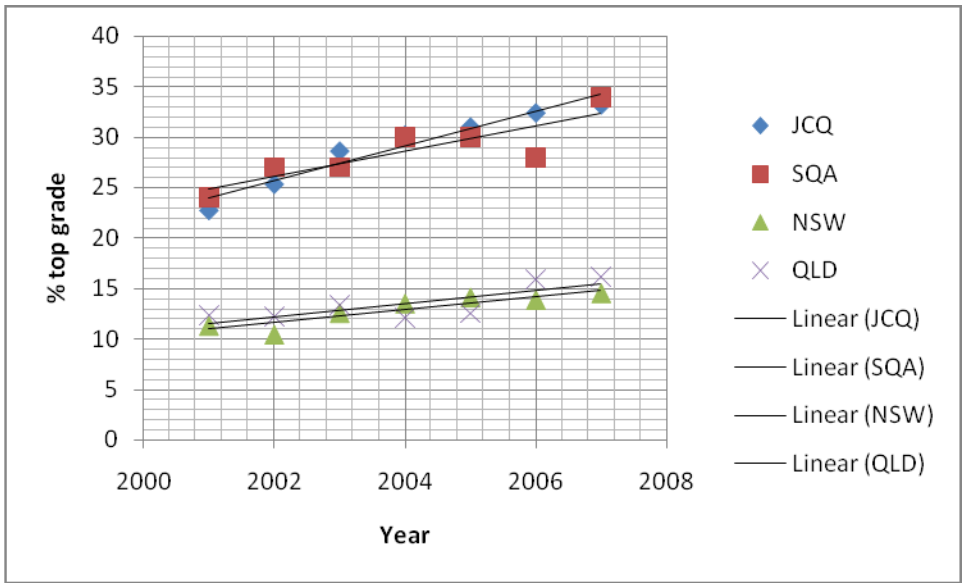


Figure 9: Top percentage for Economics

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.

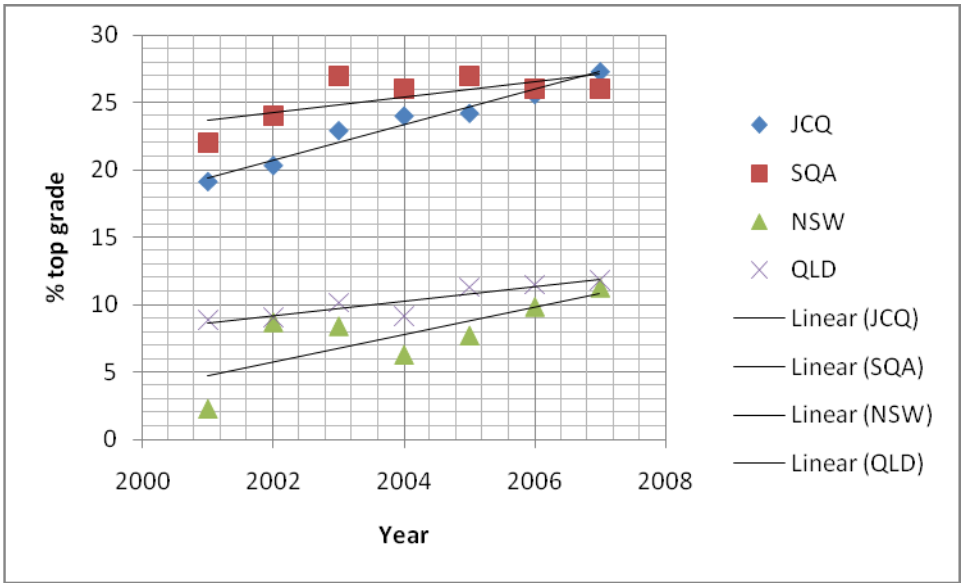


Figure 10: Top percentage for Geography

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

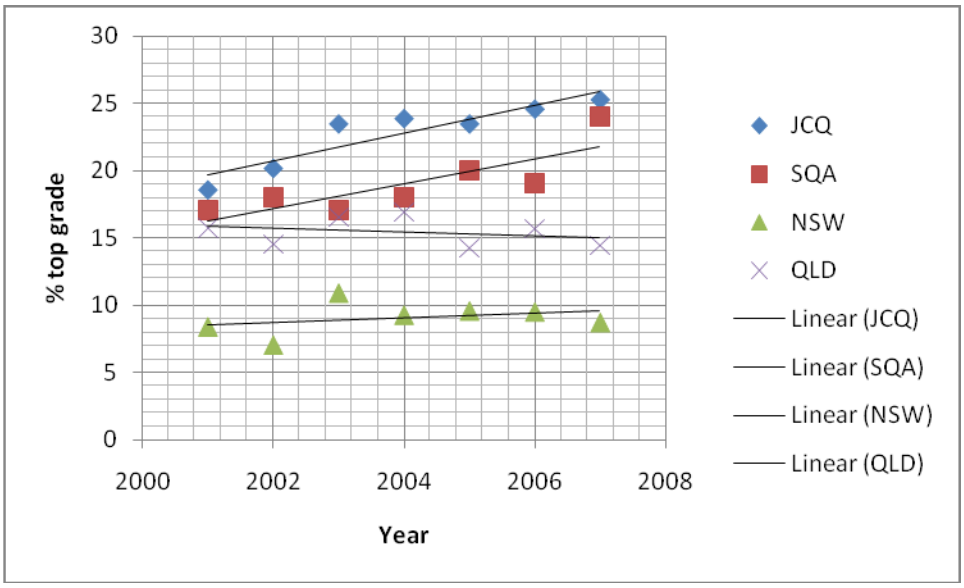


Figure 11: Top percentage for History

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

Email

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

How Can We Validate Educational Outcomes When Reported In Terms Of Standards?

Jim Tognolini
Senior Vice President, Pearson (Research and Assessment) and Senior Research Fellow at Oxford University

Gordon Stanley
Honorary Professor, Sydney University and Senior research Fellow at Oxford University

ABSTRACT

Most educational systems have moved from a norm-referenced ('grading on a curve') to a standards-referenced system of reporting educational outcomes. Instead of a fixed proportion (e.g 10%) of a cohort being assigned the top grade, the latter type of reporting requires judgments to be made about where to place the cut-score on a distribution of marks to indicate achievement of the required grade standard for each grade level awarded. A grade is only given to those students who have demonstrated the criteria for the grade. Such reporting makes sense when the intention is to interpret student outcomes in terms of explicit standards. The potential down-side of such reporting is that there may be subtle and not so subtle pressure on the judgment process to inflate student achievement. Grade inflation appears to be occurring in a number of education systems and seems to be an unfortunate potential by-product of standards-referenced reporting. In this paper the authors discuss quality assurance processes and measures needed to validate whether changes in the distribution of results with standards-based reporting of grades is real or inflated.

Introduction

Education systems around the world have been moving from norm-referencing to standards-referencing their reporting of educational outcomes (Tognolini & Stanley, 2007). One of the main differences between norm- and standards-referencing is that with the latter there is no inherent limit to the percentage of students achieving a particular standard. In theory it is possible, though unlikely, for all students to achieve any performance standard. This opens up the possibility of grade inflation occurring.

Grade inflation refers to the situation where grades appear to be improving over time without any corresponding evidence suggesting improvement. If over time the characteristics of the students presenting for the examination are not changing and there is no evidence of any change in the teaching/learning process there would be suspicion about the validity of grade increases.

In the UK, higher education institutions typically use a standards-based reporting system. Expansion of higher education was made on the assumption that common standards are being used in reporting student outcomes. However data on student performance has raised issues about grade inflation. There is skepticism about the rate of upper seconds and firsts being given in degree results which has shown an almost 8 percent increase from 1994-2007. Yorke et al (2002) found that 22% of UK first degree awards in Mathematics were at first class level, while for law it was only 4%. They concluded that this variation appeared to have little relationship at all to any identifiable measure of input.

In standards-referenced systems the percentage achieving particular performance bands or levels can vary from year to year. The question is how can stakeholders know that the percentage reported as achieving the bands is derived from a comparable set of information from year to year? Is it good enough to just attest that due process has been carried out or is there a need for more substantive information regarding the percentages produced?

The emphasis in these questions is whether the alignment of the cut-score to the 'borderline student' from one year is equivalent to the new cut-score of the same borderline student in subsequent years. Given that there is always a degree of uncertainty at the decision point for a cut-score, the tendency for slight movements in one direction may have little effect (1-2%) for a given cohort's performance. However if every year there is a small 'downward' shift in the cut-score the cumulative effect over several years can lead to major grade inflation as occurred in the English A levels (de Waal & Cohen, 2007).

Clearly this issue is of considerable interest when there is pressure on school systems to demonstrate that outcomes are improving or not moving backwards. Assessment authorities need to take the issue of validating the reported outcomes seriously. Of course in so doing consideration has to be given to the cost efficiency of such procedures.

It is important to create alternate multiple sources of information that indicate the relative stability of the results of the standard setting exercise. If these different sources give similar information (convergent validation) then authorities can be more confident that the results are comparable and any change is genuinely a change in the distribution of performance from one year to the next.

Methods

Options for collecting validating information

Assume that for the current year a professional judgment-based standard setting exercise has been conducted by the examination board and the percentages achieving particular grade levels determined. The judges would have been required to set cut scores on mark distributions using one of the common standards setting procedures such as Angoff (1971) or book-marking (Mitzel et al, 2001). Judges typically are drawn from experienced teachers and subject specialists who are assumed to have a clear understanding of the standard to which student work is to be referenced. For them the standard would have become internalised through experience with student work in a number of contexts.

The distribution of performance from a standards-referenced system using professional judgment procedures does not automatically deliver an identical distribution for each subject each year because the results delivered by the judges are not aligned to a pre-determined normal distribution as occurs with 'grading on a curve'. Any change in distribution should be indicating something about the real performance characteristics of the current cohort, and not be an artifact of the judgment process itself.

Thus there is a lot at stake for ensuring that the standards setting process is not captured by good intentions with respect to 'giving the benefit of any doubt' about where to place the cut-scores to the current cohort of students. As mentioned earlier such generosity of spirit by judges can result in small drifts over time perverting the course of valid standards-setting and lead to grade inflation. Judges need to be trained to resist such tendencies but collection of additional data can be useful to check that such influences are not at work.

If the current distribution of marks/grades is different from the previous year how can one be confident that the judgment process has delivered a valid outcome? What are some 'other' ways that alternate sources of information about the current distribution can be obtained (relative to the previous distributions) that will enable the validation of the outcomes of the results of the current standard setting exercise?

In order to be able to compare one distribution to another there is a need to be confident that they are resulting from marks being on a common scale. For this to be achieved it is necessary to have something in common. It could be common items (items, questions, tasks, examinations or tests), a common generic test (sometimes called a moderating test), common students (i.e. students who do both examinations and don't change between the first and second, an unlikely proposition if the examinations are a year apart) and/or common judges (i.e. judges internalise the standards which informs their professional judgement).

Statistical Equating and Moderation

With public examinations it is rare that papers are kept secure; so re-using all the same items or questions is not a likely option. One solution used by some systems is to have a set of common items from both years tests embedded in a form given to a population from another jurisdiction so that some statistical equating of the difficulty of each year's test can be made. Students from a similar, but different, system are asked to complete a shortened composite paper that comprises items (that assess material that is known to the students in the chosen system) from the years that need to be equated or compared.

The results can then be used to place the distributions from the current year cohort onto a common scale so that the cut-scores across the current and previous year can then be compared. Clearly this option is not always possible and is problematic if there are significant differences in curriculum across the two jurisdictions. Moreover under such circumstances it is difficult to achieve the degree of motivation characteristic of a 'live' examination.

Another approach for aligning performances from different distributions onto a common scale can be achieved by using a generic moderating test (Core Skills or General Achievement Test). This test can be administered to the whole cohort or to a sample of students in a sample of subjects each year. Such tests need to be kept secure. The distributions of results from different years can then be mapped onto the scale of the moderating test and comparisons can then be made to make sure that the cut-scores do align (within reason).

This approach makes most sense with academic subjects where it can be assumed that there is a common academic ability underlying performance outcomes. With a general aptitude test a check can be made as to whether or not the general ability level of the cohort of students in the current year is different from previous years. General aptitude tests administered to whole candidatures are common practice in some Australian state public examination systems (Queensland, ACT and Victoria). If there is no change in the general ability profile of the candidature one might question any improvement in the distribution of performance in any particular academic subject unless there is other evidence to confirm it.

If calibrated item banks are used to develop the moderating tests then the security of the moderating tests is not a major issue. As with using common items with common students from another jurisdiction to create a common scale, there are some advantages and disadvantages with the use of a moderating test approach as shown in Table 1.

Table 1
Advantages and disadvantages of moderating tests

	Advantages		Disadvantages
1.	It is perceived to be an alternate to professional judgement	1.	Relatively costly and quite intrusive
2.	It is well known and accepted as a method to equate and compare distributions	2.	May be difficult to motivate students ~ this could lead to a diminution of validity
3.	One single test can be used to accommodate most subjects and sub-tests of the test can be used to equate the different subjects	3.	Generic tests are only loosely linked to the actual content in the examinations
4.	Actual student performance is used to compare the subjects	4.	Adds to the examination load of students
		5.	Statistical in nature and would be relatively difficult for teachers and the community to understand
		6.	Security is an issue

Audit by Using Additional Professional Judgment

In addition to the initial panel of judges an audit or verification panel could be used to interrogate the data and process to make their own independent, professional judgment about the relative differences between the distributions from the different years. This could involve interviewing the examiners, markers, judges and asking them such questions as “Is this year’s paper more difficult than last year’s”; “Is there a difference in the ability of this year’s cohort relative to the previous year?”; etc.

Should feedback from these audit questions suggest that the examination was perceived to be less difficult than that of the previous year, but that judges had set lower cut-offs, then there would be prima facie evidence that the judges were being lenient. Instead of correcting their cut-scores upwards to take account of the easier paper, they have moved their cut-scores in the opposite direction.

If their original cut-scores were allowed to stand this would lead to an inflated result for the current year relative to the previous year. The authors have observed such an outcome, the inconsistency not being recognised by members of the original judge panel until the audit questions were asked. Ideally one would hope that the audit process would not reveal any inconsistencies which lead to doubt about the current cut -scores being consciously or unconsciously ‘gamed’.

Of course the perception of a paper being 'easier' is itself a judgment that may not be evidence-based. Judges may be influenced by feedback from students who may be better prepared and hence find the paper easier than they expected. To correct for such performance would be to over-ride genuine improvement.

Clearly examination authorities need to be aware of these possibilities and ensure enough evidence is obtained to resolve what might otherwise be distorting influences in finalising cut-scores.

Supplementary Judge Panels

Another approach to validation would be to have a completely independent standard setting exercise using equivalent panels of judges. Depending on the size of the local education community having two independent panels for each subject domain from within the same school system may not be possible. An alternate strategy would be to use judges from a different educational system who are familiar with the curriculum in the original system.

There are some practical limitations to implementing parallel panels of judges. Setting up panels of judges and running standards setting exercises is logistically quite a task if the examination authority is responsible for assessments across a wide range of subject domains. Expense and organisational demands including time-critical decisions makes this a less likely option.

Another professional judgment approach to providing some validating information is to ask the examiners to estimate the cut-scores when they set the examination. This would enable a comparison of the 'intended' cut-scores with those obtained by the panel of judges in the standards-setting exercise. This process would be fairly simple to implement and is already common practice is some systems.

Typically where the examinations are high stakes the previous years' examinations have been used by teachers and students to prepare for the current examination. The performance standards used to assign grades are available to school systems so teachers and students have the opportunity to "internalise" the performance standards. In these contexts it would be possible to have teachers in the system also estimate the cut-scores on the examination.

The process for teachers could begin after the examination has started and before the examination is complete so that the students have not contaminated the judgement by providing their views on the relative difficulty of the paper to the teachers (see below). Teachers generally are keen to look at the examination papers and to make judgments about he fairness or otherwise of the papers. Modern technology makes it possible to have real time access to papers on-line and teachers could access a secure site to participate.

Both the examiner and teacher estimates of current year cut-scores could work as described when the paper/item difficulty is relatively easily seen by inspection of the paper as tends to be the case in mathematics and science oriented subjects. It would be more problematic in those subjects in the humanities area when the questions might be quite general and 'accessible' and student answers need to be seen to identify whether or not the intended comparability of papers occurred.

In the context of public examination standards-setting it is common to adopt a multistage process in which sample scripts and item performance statistics are provided to assist the judges in selecting their cut-scores (Berk, 1966; MacCann & Stanley, 2004; Popham, 1978). This would need more than just a priori judgments.

The a priori estimates by either examiners or teachers could feed into the first stage of a traditional Angoff procedure and indicate any divergence of views for the current year relative to previous agreement at this stage in past years.

In principle using information from teachers is valuable in ensuring that the process of professional judgment is not too removed from the experience of the classroom. However as shown in Table 2 it is important to recognise both the advantages and disadvantages of using teacher judges in standards-setting.

Table 2
Advantages and disadvantages of teacher involvement as judges

	Advantages		Disadvantages
1.	Involves teachers in applying the standards; helps internalise the standards across the system	1.	It validates professional judgement with professional judgement
2.	It gives the system level authorities feedback as to how well the standard is effectively embedded	2.	Not getting student comparison only getting teacher estimates i.e. teacher effect
3.	Not statistical; relies on professional judgement	3.	Needs to be done online or by phone
4.	Relatively cheap and non-intrusive	4.	Could lack authenticity within the community because the teachers themselves are making the judgements

Online Participation by Teachers

Of course with Internet access now commonplace, it is feasible set up an online system where large numbers of teachers could log in and participate in the standards setting process. In this way a large and more represenatative sample of judges would be utilised. Any materials required for the judging could be made available online (for example, the examination paper and the marking rubric). It would be possible to deliver a training package online, using past examples, thus allowing teachers to practice rating items and recommending cut scores.

An online system could provide feedback to teachers so they could see how consistent their judgments are with the central trend of their peers and contribute to a broad embedding of knowledge about the explicit standards underlying reporting in the education system in which they are working.

The relative merits of making more use of teachers in the process as some verification of expert panel judgments or as an alternative to small expert panels need to be considered. MacCann and Stanley (2010) report system level data showing more stable judgments from a large pool of teachers than from a small expert group.

In large scale tests typically 15-30 judges is seen as desirable (Hambleton & Pitoniak, 2006). Cost considerations in traditional face-to-face small group meetings means that examination boards often find themselves working at the low end of that range. An online system once developed with appropriate security checks built in would require considerable development costs upfront. Once available it would enable large participation of teachers at marginal cost and overcome some of the problems of the representativeness of small samples of judges.

Whether or not an online system would be more prone to grade inflation and gaming strategies than existing approaches is hard to tell in advance. Such a process would enable tracking of outliers as judgments could be monitored in real time.

Combination of Teachers and Statistics

Another method for checking the consistency of cut-scores from year to year is to create a composite examination that comprises different questions from previous years examinations. Panels of examiners, judges and teachers can then be invited to take pairs of questions and compare them in terms of their relative difficulty ("pair-wise" comparisons). The data from these professional judgements can then be used in conjunction with Item response Theory (IRT) to locate the items along a single measurement continuum. The items from the various years' examinations can then be used to align the distributions across the various calendar years onto a common scale. This would enable the cut-scores to be directly compared; any variations in these scores would indicate that the cut-scores are not equivalent and that direct comparisons of the percentages achieving the various grades across time are problematic.

Conclusion

Standards-referenced reporting is important for providing information about the content of student achievement. However the process of alignment of student work to the standard involves professional judgment. The consistency of the process needs to be validated. Validation requires use of additional information to confirm the standards-setting. This paper describes some options of using moderating tests, replication of judgments and audit processes to validate the results. Online operation of standards-setting has the potential to provide for larger involvement in the process and greater stability, even if initial development costs are high.

References

Angoff W.H., 1971. Scales, norms and equivalent scores. In Educational Measurement (Ed. R.L. Thorndike) pp:508-600. Washington, D.C.: American Council on Education.

Berk, R., 1996. Standard setting: the next generation. Applied Measurement in Education, 9: 215-235.

de Waal, A. and Cowen, N. 2007. The results generation. Civitas, 08.
www.civitas.org.uk/pdf/resultsgeneration.pdf

Hambleton, R. K., and Pitoniak, M. J., 2006. Setting performance standards. In: Educational Measurement, (Ed . R.L. Brennan) pp: 433-470). Washington, DC: American Council on Education.

MacCann, R.G. & Stanley, G., 2010. Extending participation in standard setting: an online judging proposal. Educational Assessment, Evaluation and Accountability, 22: 139-157.

MacCann, R.G. and Stanley, G., 2004. Estimating the standard error of the judging in a modified-Angoff standards setting procedure. Practical Assessment Research and Evaluation, 9(5):Retrieved 1 July, 2008 from http://pareonline net/getvn.asp?v=9&n=5

Mitzel, H. C., Lewis, D. M., Patz, R. J. and Green, D. R., 2001. The bookmark procedure: psychological perspectives. In: Setting performance standards (Ed. R.L. Brennan) pp. 249-281. Mahwah, NJ: Lawrence Erlbaum.

Popham, W., 1978. As always provocative. Journal of Educational Measurement, 15: 297-300.

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.

Yorke, M., Barnett, P., Bridges,P., Evanson,P., Haines,C., Jenkins,D., Knight., P., Scurry,D., Stowell, M., & Woolf., H., 2002. Does grading method influence honour degree classification. Assessment and Evaluation in Higher Education, 27(3): 269-279.