CAMBRIDGE ASSESSMENT

Prediction matrices, choice and grade inflation

Tom Bramley

Cambridge Assessment Research Report

24th October 2013

University of Cambridge
Local Examinations Syndicate

Author contact details:

ARD Research Division
Cambridge Assessment
1 Regent Street
Cambridge
CB2 1GG

Bramley.t@cambridgeassessment.org.uk

http://www.cambridgeassessment.org.uk/

How to cite this publication:

Bramley, T. (2013). *Prediction matrices, choice and grade inflation*. Cambridge Assessment Research Report. Cambridge, UK: Cambridge Assessment.

Introduction

The ‘comparable outcomes’ approach (Ofqual, undated) to maintaining standards has become the dominant approach used by awarding organisations (AOs) when deciding where to set the grade boundaries in a given examination. This short study used simulated data to explore some of the consequences of using this approach in a hypothetical scenario. For simplicity, the scenario involved one exam, for the sake of concreteness called ‘Maths’ and two alternative exams, called ‘Linear’ and ‘Modular’ – referring to how they might be assessed.

The simulated data consisted of four normally distributed standardised variables, corresponding to ‘KS2 score’, ‘Maths score’, ‘Linear score’ and ‘Modular score’. The KS2 (Key Stage 2) score is the measure of prior attainment used in prediction matrices at GCSE. The intercorrelations of these simulated variables are shown below in Table 1. Each exam was simulated to be equally correlated (0.7) with prior attainment, the two alternatives equally correlated with the original (0.95), but slightly less well correlated with each other (0.9).

Table 1: Intercorrelations of the simulated variables (N=10,000)

	KS2	Maths	Linear	Modular
KS2	1	0.70	0.70	0.70
Maths		1	0.95	0.95
Linear			1	0.90
Modular				1

In practice we could only ever know the joint distribution of scores on KS2 and one of the three exams. In the simulation we know the joint distribution of scores on KS2 and each of the three other exams for the whole cohort. I assumed that a cohort-referenced definition of comparability (e.g. Wiliam, 1996) would be reasonable in this scenario and hence applied exactly the same grade distribution[1] to each of the three exams to give the ‘true grades[2]’ (A* to U) in those exams. First, the cross-tabulation of KS2 decile[3] with Maths grade (see Table 2) was used to produce the ‘prediction matrix’ used for all analyses in this report.

Table 2: Prediction matrix (cross-tabulation of KS2 decile with Maths grade) showing cumulative proportions within each decile obtaining each grade.

ks2decile	A*	A	B	C	D	E	F	G	U
1	0.32	0.619	0.835	0.975	0.997	0.999	1	1	1
2	0.106	0.337	0.619	0.903	0.977	0.993	1	1	1
3	0.056	0.226	0.474	0.834	0.956	0.993	1	1	1
4	0.036	0.158	0.381	0.755	0.927	0.972	0.991	1	1
5	0.015	0.094	0.255	0.67	0.879	0.956	0.988	0.999	1
6	0.011	0.066	0.203	0.547	0.792	0.917	0.977	0.998	1
7	0.002	0.029	0.117	0.445	0.722	0.873	0.962	0.994	1
8	0.001	0.015	0.087	0.376	0.641	0.815	0.927	0.986	1
9	0.003	0.006	0.042	0.268	0.514	0.716	0.881	0.967	1
10	0	0	0.007	0.097	0.255	0.436	0.664	0.876	1
All	5.5%	15.5%	30.2%	58.7%	76.6%	86.7%	93.9%	98.2%	100%

[1] Using the values from the overall total distribution of grades in syllabuses classified as GCSE Mathematics in the Inter-board statistics for June 2012.

[2] These are not ‘true grades’ in the sense of classical test theory, but rather the grades that would be obtained with a cohort-referencing approach in the hypothetical situation where the whole cohort took all three exams.

[3] In practice KS2 scores are split into 8 categories based on average levels across the three subjects English, Maths and Science for the purpose of producing prediction matrices.

Scenario 1. Predicting grades on the same exam for subsets of the cohort

The whole purpose of prediction matrices is to adjust for differences in the prior attainment of different exam cohorts. To illustrate how they achieve this, Tables 3a and 3b show the predicted (by the prediction matrix) and 'true' grade distributions for two non-overlapping 'high ability' and 'low ability' cohorts of 5,000 examinees. The high ability cohort was created by selecting a stratified random sample without replacement using a sampling proportion of 0.15 for the top five deciles and 0.05 for the bottom five deciles. The low ability sample comprised the examinees not selected for the high ability sample.

Table 3a: Predicted and true grades in Maths for high ability sample (N=5,000)

	%			Cumulative %		
Maths	Predicted	True	Diff.	Predicted	True	Diff.
A*	8.08	7.78	+0.30	8.08	7.78	+0.30
A	14.01	14.18	-0.17	22.09	21.96	+0.13
B	18.65	18.90	-0.25	40.74	40.86	-0.12
C	29.98	29.42	+0.56	70.72	70.28	+0.44
D	14.94	15.36	-0.42	85.66	85.64	+0.02
E	6.82	6.50	+0.32	92.48	92.14	+0.34
F	4.26	4.70	-0.44	96.74	96.84	-0.10
G	2.35	2.34	+0.01	99.09	99.18	-0.09
U	0.91	0.82	+0.09	100.00	100.00	

Table 3b: Predicted and true grades in maths for low ability sample (N=5,000)

	%			Cumulative %		
Maths	Predicted	True	Diff.	Predicted	True	Diff.
A*	2.92	3.22	-0.30	2.92	3.22	-0.30
A	5.99	5.82	+0.17	8.91	9.04	-0.13
B	10.75	10.50	+0.25	19.66	19.54	+0.12
C	27.02	27.58	-0.56	46.68	47.12	-0.44
D	20.86	20.44	+0.42	67.54	67.56	-0.02
E	13.38	13.70	-0.32	80.92	81.26	-0.34
F	10.14	9.70	+0.44	91.06	90.96	+0.10
G	6.25	6.26	-0.01	97.31	97.22	+0.09
U	2.69	2.78	-0.09	100.00	100.00	

It can be seen that the predicted distributions do indeed compensate effectively for the differences in ability with all discrepancies between true and predicted cumulative percentages being less than half a percentage point[4]. The discrepancies arise from what Benton & Lin (2011) referred to as 'model standard errors' – in that even with the correct matrix at population level, sampling fluctuations within a particular cohort will not reproduce that matrix exactly.

[4] This simulation was not designed to shed light on what the 'tolerances' for deviations from predictions should be. This is a complex matter requiring (among other things) consideration of the clustering of pupils within schools. See Benton & Lin (2011).

Scenario 2. Predicting grades when examinees have a choice of exam

Tables 4a and 4b below show the predicted and true grade distributions for a scenario where examinees choose either Linear or Modular based on perfect knowledge of which they would do better in. That is, two cohorts were created by choosing those who had got a better simulated score on Linear than Modular and vice versa.

Table 4a: Predicted and true grades in Linear for 'linear better' sample (N=5,026)

	%			Cumulative %		
Maths	Predicted	True	Diff.	Predicted	True	Diff.
A*	5.59	7.78	-2.19	5.59	7.78	-2.19
A	10.10	12.18	-2.08	15.69	19.96	-4.27
B	14.76	17.05	-2.29	30.45	37.01	-6.56
C	28.36	29.23	-0.88	58.81	66.24	-7.43
D	17.79	16.37	+1.42	76.60	82.61	-6.01
E	10.05	8.36	+1.69	86.64	90.97	-4.33
F	7.19	5.37	+1.82	93.84	96.34	-2.51
G	4.34	2.65	+1.69	98.17	98.99	-0.82
U	1.83	1.01	+0.82	100	100	0.00

Table 4b: Predicted and true grades in Modular for 'modular better' sample (N=4,974)

	%			Cumulative %		
Maths	Predicted	True	Diff.	Predicted	True	Diff.
A*	5.41	7.36	-1.95	5.41	7.36	-1.95
A	9.90	11.76	-1.86	15.31	19.12	-3.81
B	14.64	16.71	-2.07	29.94	35.83	-5.89
C	28.65	29.35	-0.70	58.59	65.18	-6.59
D	18.01	16.52	+1.49	76.60	81.70	-5.10
E	10.16	8.59	+1.57	86.76	90.29	-3.53
F	7.21	5.69	+1.52	93.97	95.98	-2.01
G	4.27	3.08	+1.19	98.23	99.06	-0.83
U	1.77	0.94	+0.83	100	100	0.00

The two cohorts had very similar prior attainment, as seen by the similarity of the predicted percentages. However, the predictions were substantially below the true percentages in both cases with a maximum discrepancy of around 7 percentage points at cumulative grade C.

Would it therefore be wrong to use the prediction matrix in this scenario? On the one hand, examinees would receive a worse grade than their true grade in a significant number of cases. On the other, if the 'true grades' were to be awarded (for example if expert judges were somehow able to perceive the boundaries that would produce the true grades), then there would be substantial 'grade inflation' in the sense of the combined grade distribution across Linear and Modular being much higher (i.e. containing higher cumulative percentages) than the original Maths grade distribution.

Of course this is an extreme example because in practice examinees would not know which exam they would do better in – but they might still have some idea (as might their teachers) based on preferences, course content, assessment structure etc.

In short, this example illustrates how the comparable outcomes approach can prevent grade inflation, but raises the question of whether it is always desirable to do so. Conversely, the example also illustrates how allowing choice can create grade inflation.

Scenario 3a. Predicting grades when examinees can switch exams mid-course

A concern in recent years with modular GCSEs has been that some candidates are switching from modular to linear syllabuses part way through the course, on the presumed basis that poor scores on units taken early mean that candidates cannot recover sufficiently on the later units to achieve their desired grades, unless they re-sit the earlier units. Because the entire linear examination is taken at the end of the course, an examinee can redeem themself by extra learning (or benefit from good luck).

The example in Tables 5a and b was created by assuming that initially examinees would choose Modular or Linear at random, that 10% of examinees would switch from Modular to Linear, and that the probability of switching would depend on the true grade in the Modular. The triangular distribution[5] was used to create the probability of switching, with a peak in the middle of the grade D range of scores, a lower limit at the bottom of the ability range and an upper limit towards the lower end of the grade C range of scores. This was to embody the realistic assumption that those examinees heading for a grade D are most likely to switch (or be switched by their schools). The sample sizes were chosen such that the numbers taking Modular and Linear would be equal after examinees had switched.

Table 5a: Predicted and true grades in Linear (N=5,000) including switchers from Modular.

	%			Cumulative %		
Maths	Predicted	True	Diff.	Predicted	True	Diff.
A*	4.98	5.14	-0.16	4.98	5.14	-0.16
A	9.38	8.46	+0.92	14.36	13.60	+0.76
B	14.19	12.92	+1.27	28.56	26.52	+2.04
C	28.47	28.18	+0.29	57.03	54.70	+2.33
D	18.33	18.58	-0.25	75.36	73.28	+2.08
E	10.52	11.70	-1.18	85.89	84.98	+0.91
F	7.59	8.04	-0.45	93.48	93.02	+0.46
G	4.58	4.98	-0.40	98.06	98.00	+0.06
U	1.94	2.00	-0.06	100	100	0.00

Table 5b: Predicted and true grades in Modular (N=5,000) without switchers to Linear.

	%			Cumulative %		
Maths	Predicted	True	Diff.	Predicted	True	Diff.
A*	6.02	6.26	-0.24	6.02	6.26	-0.24
A	10.62	11.32	-0.70	16.64	17.58	-0.94
B	15.21	16.44	-1.23	31.84	34.02	-2.18
C	28.53	29.74	-1.21	60.37	63.76	-3.39
D	17.47	15.76	+1.71	77.84	79.52	-1.68
E	9.68	8.42	+1.26	87.52	87.94	-0.42
F	6.81	6.72	+0.09	94.32	94.66	-0.34
G	4.02	3.78	+0.24	98.34	98.44	-0.10
U	1.66	1.56	+0.10	100	100	0.00

It can be seen that the effect of the switch was for the prediction matrix to predict slightly better grades than the true grades for the Linear exam, and slightly worse grades for the Modular exam, with the most noticeable impact at cumulative grade C. The explanation for this can be found by considering Table 6.

[5] See for example http://en.wikipedia.org/wiki/Triangular_distribution

Table 6: Mean and SD of scores on each test before and after switching

		KS2		Maths		Linear		Modular	
	N	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Cohort									
Linear initial	4444	0.00	1.00	-0.00	1.00	-0.00	1.01	-0.00	1.00
Modular initial	5556	-0.00	1.00	0.00	1.00	0.00	0.99	0.00	1.00
Switchers	556	-0.58	0.81	-0.82	0.58	-0.77	0.62	-0.86	0.52
Linear final	5000	**-0.06**	1.00	-0.09	1.00	**-0.09**	1.00	-0.10	0.99
Modular final	5000	**0.06**	1.00	0.09	0.99	0.09	0.99	**0.10**	1.00

The switchers from Modular had lower simulated scores, on average, in the Modular exam than on the KS2 exam. This is a direct consequence of the 'regression to the mean' built into the simulation (linearly related scores with a correlation of 0.7[6]). Because the prediction matrix only considers KS2 performance it therefore overestimated the performance of the switchers, thus resulting in predictions for the switched-to exam (Linear) that were slightly higher than the true grades, and vice versa for the switched-from exam (Modular).

This overestimation was not compensated for by the fact that the switchers had slightly better (on average) scores and hence true grades on the switched-to than the switched-from exam. That is, switchers below average on Modular had (on average) slightly higher scores on Linear but because the correlation between the two was higher than with KS2 there was less regression to the mean. In this scenario (unlike the previous one) the relative performance on Modular and Linear played no part in the selection of switchers.

Scenario 4b. Predicting grades when examinees make an informed switch mid-course

Repeating scenario 4a (10% switching from Modular to Linear) but this time assuming 'perfect knowledge' on the part of the switchers by using the same triangular distribution but setting to zero the probability of switching for those with higher scores on Modular than Linear gave the results in Tables 7a and b below.

Table 7a: Predicted and true grades in Linear (N=5,000) including switchers from Modular who all had higher scores in Linear than Modular.

	%			Cumulative %		
Maths	Predicted	True	Diff.	Predicted	True	Diff.
A*	5.00	5.14	-0.14	5.00	5.14	-0.14
A	9.40	8.46	+0.94	14.40	13.60	+0.80
B	14.23	13.02	+1.21	28.63	26.62	+2.01
C	28.49	29.86	-1.37	57.12	56.48	+0.64
D	18.34	19.16	-0.82	75.45	75.64	-0.19
E	10.51	10.86	-0.35	85.97	86.50	-0.53
F	7.56	7.12	+0.44	93.53	93.62	-0.09
G	4.55	4.52	+0.03	98.08	98.14	-0.06
U	1.92	1.86	+0.06	100	100	0.00

[6] The same effect can be observed for the Maths and Linear scores of the switchers, but it is much less noticeable because the simulated correlations were much higher (0.95 and 0.9 respectively).

Table 7b: Predicted and true grades in Modular (N=5,000) without switchers to Linear who all had higher scores in Linear than Modular.

	%			Cumulative %		
Maths	Predicted	True	Diff.	Predicted	True	Diff.
A*	6.00	6.26	-0.26	6.00	6.26	-0.26
A	10.60	11.32	-0.72	16.60	17.58	-0.98
B	15.18	16.44	-1.27	31.77	34.02	-2.25
C	28.51	29.78	-1.27	60.28	63.80	-3.52
D	17.46	16.00	+1.46	77.75	79.80	-2.05
E	9.69	8.48	+1.21	87.43	88.28	-0.85
F	6.84	6.40	+0.44	94.27	94.68	-0.41
G	4.05	3.82	+0.23	98.32	98.50	-0.18
U	1.68	1.50	+0.18	100	100	0.00

Comparing Table 7a with Table 5a we see that 'informed switching' reduced the discrepancy between predicted and true grade for the Linear exam. Rather than predicted grades being higher than true grades (at all cumulative grades except A*), predicted grades were slightly higher at A to C and slightly lower at D to G, with all discrepancies apart from grade B being of similar size to those found in the high-low ability split (scenario 1). In contrast, there was much less effect on the Modular exam losing the switchers. Comparing Table 7b with Table 5b it can be seen that predicted grades were lower than true grades in both, by approximately the same amount. Table 8 shows the equivalent information to Table 6:

Table 8: Mean and SD of scores on each test before and after switching (switchers to Linear all had higher scores in Linear than Modular).

		KS2		Maths		Linear		Modular	
	N	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Cohort									
Linear initial	4444	0.00	1.00	-0.00	1.00	-0.00	1.01	-0.00	1.00
Modular initial	5556	-0.00	1.00	0.00	1.00	0.00	0.99	0.00	1.00
Switchers	556	-0.53	0.79	-0.69	0.56	-0.49	0.56	-0.88	0.53
Linear final	5000	**-0.06**	0.99	-0.08	0.99	**-0.06**	0.98	-0.10	1.00
Modular final	5000	**0.06**	1.01	0.08	1.00	0.06	1.01	**0.10**	0.99

The comparison of the cells in bold between Tables 6 and 8 shows why the predictions were closer to the true grades in the switched-to (Linear) exam than the switched-from (Modular) exam when there was informed switching. The switching resulted in the final set of Linear examinees having the same score distribution (mean and SD) on KS2 as on the Linear exam; whereas the final set of Modular examinees still had higher scores on the Modular exam than on KS2.

Summary and conclusion

The critical assumptions of the simulation of the main data set were that:

- underlying scores on all exams were normally distributed and hence linearly related;
- the correlation between prior attainment (KS2) and GCSE was less than the correlation among different GCSE exams in the same subject.

If these assumptions are justifiable then the example scenarios suggest the following implications:

If one exam is replaced by two (or by extension one board/syllabus is replaced by two etc.) then a prediction matrix based on the whole cohort in the old exam will predict lower grades than the 'true grades' in both the new exams, provided examinees have some insight into which of the two new exams they would score relatively higher in.

If below average examinees switch from one exam to another[7] then then a prediction matrix (based on a different exam that is equally correlated with both) will over-predict true grades on the switched-to exam and under-predict true grades on the switched-from exam, assuming switchers are basing their decision to switch purely on how they perceive their performance on the switched-from exam.

If below average examinees switch from one exam to another with some insight into whether they will score higher in the switched-to than the switched-from exam then the discrepancy between predicted and true grades will reduce for the switched-to exam, but not for the switched-from exam.

It is a matter for debate whether the desirable outcome is for the awarding process to award the true grades according to the original cohort-referenced definition of true grades. If it is, then grade inflation could be a consequence when examinees are given more choice of exams.

However, it would be inconsistent to argue that the awarding process should aim to produce the predicted grade distribution rather than the true grade distribution when examinees are given more choice (thus preventing grade inflation) but also argue that the awarding process should aim to award the true grade distribution rather than the predicted grade distribution when examinees switch from one exam to another (thus potentially allowing grade inflation).

References

Benton, T., & Lin, Y. (2011). Investigating the relationship between A level results and prior attainment at GCSE. Coventry: Ofqual.

Ofqual (undated). GCSEs and A Levels in Summer 2012. Our approach to setting and maintaining standards. Coventry: Ofqual. http://www2.ofqual.gov.uk/files/2012-05-09-maintaining-standards-in-summer-2012.pdf Accessed 24/10/13.

Wiliam, D., & Black, P. (1996). Meanings and consequences: a basis for distinguishing formative and summative functions of assessment. *British Educational Research Journal, 22*(5), 537-548.

[7] The examples used the terms Modular and Linear to give a concrete and realistic context. But the data was all simulated, so says nothing about actual modular or linear examinations.