Partial absences in GCSE and AS/A level examinations

Carmen Vidal Rodeiro Research Division

Introduction

There are certain situations in which a candidate does not have a mark for a unit/component in a GCSE or AS/A level examination. For example, if they were ill on the day of the exam, if their paper was lost (e.g., at the centre (school), in the post, at the scanning bureau or at the awarding body's offices) or if their controlled assessment was invalid as a result of individual or centre malpractice.

Subject to certain rules, the awarding body can calculate an estimated mark for the unit/component with the missing mark to enable the candidate to certificate, rather than having to wait for the next assessment opportunity. The conditions under which an estimated mark can be awarded are set out by the Joint Council for Qualifications (JCQ, 2016).

There have been reports in the press (e.g., Espinoza, 2015; Linning, 2015) about awarding bodies 'guesstimating' hundreds of students' grades. However, a spokesman from the Office of Qualifications and Examinations Regulation (Ofqual) (quoted in the above press reports) said that a very small number of marks can be and are estimated each year and only in some very specific circumstances. In fact, he said, this number represents just a very small fraction of the number of overall papers marked.

The aims of the research described here were as follows:

1. To investigate the numbers of unit/component marks in GCSE and AS/A level qualifications awarded by the OCR awarding body that were estimated in a specific session.
2. To evaluate current and potential new method(s) for estimating missing marks. In particular, this research explored the use of statistical methods for handling missing data, specifically *regression imputation*, to estimate the mark for a missing unit/component in GCSE and AS/A level qualifications. The marks (and grades) obtained in this way were compared with the marks (and grades) obtained applying two different methods currently used by some of the awarding boards in England: the *z-score method* and the *percentile (cum% position) method*.

Data and methodology

In this research, unit/component level data from the OCR awarding body (June 2015 session) was used.

For the investigation of methods for estimating missing marks, the following analyses were carried out:

- **Simulation of missing data:** missing marks for a specified number of candidates were simulated in several GCSE and AS/A level units/components. Different strategies, which are described later, were used for this.
- **Estimation of the missing marks using three different methods:** Regression imputation, *z-score method* and *percentile method*.

Creation of missing marks

Several OCR qualifications, both at GCSE and AS/A level, with different structures (e.g., different number of units; different types of assessment) were selected for analysis. Table 1 (on page 24) gives details of the specifications included in this work.

In each of the specifications listed in Table 1, units were selected as shown in Table 2 (on page 24), and missing marks were then generated.

Partial absences for candidates certificating in June 2015 in OCR qualifications (GCSE and AS/A levels) were examined to give an idea of the numbers of candidates who are issued estimated grades in a given session. There were 19 GCSE units/components with at least 40 candidates with missing marks, and 11 units with at least 60. At AS/A level, there were several units with more than 40 candidates who had estimated marks but just 1 with more than 50. Taking this information into consideration, it seemed reasonable to select, in each unit/component listed in Table 2, 60 candidates to create missing marks for.

The different strategies to create the missing marks were as follows:

1. Candidates were selected at random and their marks in the unit/component of interest were set to missing.
2. The probability of having an absent mark for the unit/component of interest was modelled, using a logistic regression, as a function of the overall qualification grade:

   $$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 Grade_i + e_i$$

   where p_i is the probability of candidate i being absent and *Grade* is the overall qualification grade.

   This probability was used to ensure that, if certain grades were more prominent amongst candidates with partial absences, this was reflected in the sample.
3. Using unit/component level data from OCR qualifications, candidates with missing marks in June 2015 (due to missing scripts or special consideration) in any unit/component were selected. This was done for GCSE and AS/A level units/components separately.

   Candidates in the unit/component of interest with a missing mark in any other unit/component (in a qualification of the same level) had their mark set to missing. If there were more than 60 candidates fulfilling this condition, 60 were selected at random amongst them.

Methods to estimate missing marks

Z-score method

In order to estimate a missing mark when a candidate is absent from an examination in a specification which uses uniform marks (most unreformed GCSE and AS/A levels), most of the JCQ awarding bodies have been employing the same procedure, known as the *z-score method*.

Table 1: OCR qualifications considered in the analyses, June 2015

Qualification	*Subject*	*OCR specification*	*Unit/Component*	*Type of assessment*	*Weighting*	*Maximum UMS[a] marks*
AS	Biology	H021	F211	written paper	30%	90
			F212	written paper	50%	150
			F213	coursework	20%	60
AS	Media Studies	H140[b]	G321	coursework	50%	100
			G322	written paper	50%	100
			G323	written paper	50%	100
GCSE	Business and Communication Systems	J230	A265	written paper	50%	120
			A266	controlled assessment	25%	60
			A267	practical examination	25%	60
GCSE	Religious Studies B	J621	B601	written paper	25%	50
			B602	written paper	25%	50
			B603	written paper	25%	50
			B604	written paper	25%	50
GCSE	Mathematics B (linear)	J567[c]	01 (F)	written paper	50%	100
			02 (F)	written paper	50%	100
			03 (H)	written paper	50%	100
			04 (H)	written paper	50%	100

a. Uniform Mark Scale used for modular assessments. For the Mathematics B (linear) J567, this column shows the maximum possible raw mark.
b. To obtain an AS level in Media Studies candidates needed to take unit G321 and either unit G322 or unit G323.
c. To obtain a GCSE in Mathematics B candidates needed to take components 01 and 02 (Foundation Tier [F]) or 03 and 04 (Higher Tier [H]).

Table 2: Units/Components for which missing marks were generated, June 2015

Unit/ Component	*Qualification/Subject*	*Type*
F212	AS Biology	written paper
F213	AS Biology	coursework
G322	AS Media Studies	written paper
A267	GCSE Business and Communication Systems	practical examination
B601	GCSE Religious Studies B	written paper
J567/01	GCSE Mathematics B (linear)	written paper

Under this procedure, the difference between the candidate's estimated mark and the performance of candidates generally on the unit in question is the same as the average difference between the candidate's performance and the performance of candidates generally on the other units.

If the candidate performed on average slightly better (or worse) than candidates generally in other units, then the estimate for the missing mark will be slightly above (or below) the general performance on the unit of interest. The difference between the performance of the candidate in question and the performance of candidates generally is measured in terms of standard deviations. The number of standard deviations above or below the mean is called the *z-score*.

An illustration of how the method works is given below.

Example

In a three-unit specification, the average uniform marks and the standard deviation for all candidates on Units 1, 2 and 3 are given in Table 3 below. Unit 1 accounts for 30% of the assessment, Unit 2 for 50% and Unit 3 for 20%.

Let us assume a candidate scores 58 on Unit 1 and 104 on Unit 2, but is absent for Unit 3. The estimated mark for Unit 3 is calculated as shown below.

Table 3: Candidate's and average performance

	Weighting	*Average uniform marks*	*Standard deviation*	*Candidate's mark*
Unit 1	30%	50	8	58
Unit 2	50%	80	12	104
Unit 3	20%	38	3	*absent*

The candidate's mark on Unit 1 is one standard deviation (8 marks) above the average for all candidates in that unit. The candidate's mark on Unit 2 is two standard deviations (24 marks) above the average for that unit. So, taking into account the weightings of the units, the average of the standard deviations is:

$$\frac{30 \times 1 + 50 \times 2}{30 + 50} = 1.625$$

Thus, the estimated mark for Unit 3 is:

$$\text{average mark} + 1.625 \times \text{standard deviation} = 38 + 1.625 \times 3 = 42.875,$$

which is rounded to 43.

An alternative to the above method is to calculate the z-score based on the aggregation of the marks in the available units/components (candidate's total score) rather than at unit/component level. This method would have the advantage of taking into account the correlation between the marks in the different units. However, this might not be feasible in practice, as in some qualifications candidates take different optional units/components.

Percentile (cum% position) method

The principle for this method is to identify unit(s)/component(s) from the same specification for which the candidate has marks, and to award the candidate a mark for the missing unit that, as nearly as possible, places them at the same percentile of the cohort as they have achieved on the unit(s)/component(s) being used in the calculation. Where the relevant percentile occurs between two possible marks, the higher mark is awarded.

It might be recommended to align the candidate's mark in the missing unit/component with fewer units/components if there is a good reason for discounting a unit/component. For example, if a coursework unit/component is being used to estimate the mark in a written unit/component and performance is not expected to correlate much between those two units/components, it might make sense to exclude it.

An illustration of how the method works is given below.

Example

Let us assume that in a three-unit specification, a candidate is missing the mark for Unit 2, and this will be calculated using the candidate's performance in Units 1 and 3 (Table 4 below). Table 5 shows extracts of the cumulative mark distributions for Units 1 and 3.

Table 4: Candidate's performance

	Weighting	*Candidate's mark*
Unit 1	30%	72
Unit 2	50%	*absent*
Unit 3	20%	55

Table 5: Mark distributions, Units 1 and 3

	Mark on unit	*Cumulative percentage of candidates*
Unit 1	73	18.37
	72	22.12
	71	22.12
Unit 3	56	3.49
	55	7.42
	54	7.95

The cumulative percentages that correspond to marks 72 and 55 in Units 1 and 3 are 22.12 and 7.42 respectively. The next step in the method is to take the average of these two figures, taking into account the weights of the units:

$$\frac{22.12 \times 30 + 7.42 \times 20}{30 + 20} = \frac{663.6 + 148.4}{50} = \frac{812.0}{50} = 16.24$$

Looking through the mark distribution for the unit with the absent mark (Unit 2), displayed in Table 6 below, we find out the mark that corresponds to that cumulative percentage. The nearest marks on Unit 2 to one that gives 16.24 per cent of candidates are 123 and 122. A mark of 123 is hence taken.

Table 6: Mark distribution, Unit 2

Cumulative percentage of candidates	*Mark on Unit 2*
13.29	124
15.73	123
17.65	122

Regression imputation

Many missing data methods fall under the general heading of imputation. The basic idea of those methods is to substitute each missing value with some reasonable prediction (imputation). There are lots of different ways to impute missing values. In this research, *regression imputation* was used.

A regression model was fitted to predict the values of a dependent variable (marks in the unit/component of interest) based on other independent variables potentially related to the missing data (e.g., performance in other units/components of the same qualification, characteristics of the candidates). The model was then used to impute values in cases where the dependent variable was missing.

Two different regression techniques were used: *ordinary least squares* (OLS) and *quantile regression*. OLS models the conditional mean of the response or dependent variable as a function of one or more independent variables. Quantile regression models the conditional quantiles (in particular, the median), rather than the mean.

In this research, the following information for each candidate was available:

- Performance in other units/components for the same qualification.
- A measure of overall performance. This was calculated, using principal component analysis, for each candidate who had taken at least one OCR assessment in the June 2015 session. It reflects the marks achieved on all the assessments taken (excluding the score for the particular unit/component being imputed).
- Characteristics of the candidate (gender, year group, socio-economic level[1]).
- Characteristics of the school (type[2], overall attainment of its pupils[3]).

Some of the information on the candidates was obtained from the OCR awarding body and some from the National Pupil Database (NPD)[4]. Information from the NPD (e.g., socio-economic level, type of school and the overall attainment in each school) was matched to OCR data using candidate and centre numbers.

An illustration of how the method works is given below.

Example

Let us assume that a candidate sat a three-unit specification and was missing the mark in Unit 2 (as seen in Table 4 above).

A very simple linear regression model, just to illustrate the method, was fitted. The dependent variable was the mark in Unit 2 and the independent variables the marks in Unit 1 and Unit 3 (more complex models will be used in the analyses presented in this article). The fitted model is as follows:

[1] The socio-economic level was measured by the IDACI (Income Deprivation Affecting Children Index). This index measures the percentage of children in a small area around the student's home who live in families that are income deprived.

[2] Independent schools, selective schools, state-maintained schools (including comprehensive, secondary modern and academy schools), sixth form colleges and further education (FE) colleges.

[3] School average GCSE or A level performance, depending on the qualification analysed.

[4] The NPD, compiled by the Department for Education, is a longitudinal database for all children in schools in England, linking student characteristics to school and college learning aims and attainment. The NPD holds pupil and school characteristics such as gender, ethnicity or level of deprivation (IDACI) matched to pupil level attainment data.

$$\widehat{Unit\ 2} = 5.93 + 1.21\ (Unit\ 1) + 0.44\ (Unit\ 3)$$

The model above is used to predict the mark in Unit 2 for the candidate shown in Table 4:

$$5.93 + 1.21\ (72) + 0.44\ (55) = 117.25$$

The estimated mark, based on the candidate's performance in the other two units that contribute to the qualification, would be 117.

Comparison of methods

In order to compare the performance of the three methods described above, the following measures were used:

1. An indicator of how close the estimated mark was to the actual mark, over all candidates with estimated marks (n):

$$\sum_{i=1}^{n} abs\ (mark - estimated\ mark)$$

Quantile regression is designed to be the line of best fit that minimises this indicator.

2. Correlation coefficients between estimated and actual marks:

$$Corr\ (marks, estimated\ marks)$$

3. The root mean square error (RMSE) of the estimated marks:

$$\sqrt{\frac{\sum_{i=1}^{n} (mark - estimated\ mark)^2}{n}}$$

where the sum is over all candidates with estimated marks (n).

This measure is minimised by OLS regression.

These three statistics were used to compare the performance of the three methods for estimating missing marks using the candidates' marks. Qualification grades awarded based on estimated marks were also compared with actual grades, using a variety of simple descriptive statistics.

Results

Partial absences in GCSE and AS/A level units

A brief investigation into partial absences for candidates certificating in June 2015 was carried out to illustrate the numbers of candidates who were issued estimated grades. In June 2015, very small numbers of candidates received an estimated grade. The AS/A level unit with the highest number of estimated marks was F213 (Practical Skills in Biology 1), where 53 candidates out of 36,582 (0.14% of the entry) were missing the mark. At GCSE, B712 (Science modules B2, C2 and P2) had 112 candidates with an estimated mark (0.23% of the entry).

Overall, only 1,073 AS/A level missing scripts had estimated marks in June 2015, which is below 0.1% of the total number of AS/A level scripts marked by OCR. Similarly, at GCSE, 2,289 (0.08%) missing scripts had estimated marks.

In this research, instances of malpractice were not considered and only absences due to missing scripts or special consideration (the candidate was ill on the day of the exam) are included in the tables.

Estimating missing marks

This section reports on the results of the three different methods used to estimate the generated missing marks in the units/components listed in Table 2.

In the case of the regression imputation (for both the OLS and the quantile regression techniques), two different models were fitted:

(a) A model only including, as independent variables, the marks for the other units/components in the specification. In this case, the information used in the imputations is the same as the information included in the percentile and z-score methods.

Table 7: Differences between actual and estimated marks

Unit/ Component	*Missing data generation*	*Sum of absolute differences*					
		Percentile	*z-score*	*Regression imputation*			
				OLS		*Quantile regression*	
				(a)	*(b)*	*(a)*	*(b)*
F212	Scenario 1	990	964	823	746	813	742
	Scenario 2	986	963	844	661	838	659
	Scenario 3	906	903	770	613	761	608
F213	Scenario 1	420	414	349	321	349	321
	Scenario 2	482	506	386	337	378	332
	Scenario 3	463	473	485	347	482	342
G322	Scenario 1	823	810	678	634	676	629
	Scenario 2	780	752	551	561	549	562
	Scenario 3	-[a]	-	-	-	-	-
A267	Scenario 1	401	393	417	384	413	385
	Scenario 2	461	432	426	391	423	394
	Scenario 3	-	-	-	-	-	-
B601	Scenario 1	323	319	305	300	310	303
	Scenario 2	273	260	262	239	261	238
	Scenario 3	277	273	280	270	278	268
J567/01	Scenario 1	330	329	309	283	310	290
	Scenario 2	386	404	401	371	401	371
	Scenario 3	390	398	380	359	382	363

a. The '-' in the table indicates that there were no candidates taking the unit who had partial absences in any other unit at this level (GCSE or AS/A level) in the June 2015 session.

(b) A model including, as independent variables, the marks for the other units/components in the specification, information about the candidates (gender, year group), their overall performance and characteristics of the schools.

Note that model (b) has been proposed for research purposes and that there are practical limitations (e.g., data acquisition, timescales) which might mitigate against its use in an operational setting.

Table 7 shows the differences between the actual and estimated marks (absolute values) for all units/components considered in this research and in all three scenarios described in the Data and methodology section. This statistic indicates how close the estimated marks are to the actual marks (over all partial absent candidates) and, therefore, the method with the smallest value would be more desirable.

OLS and quantile regression provided very similar results for all units/components in the three different scenarios. Furthermore, and not unexpectedly, the models including additional information on the candidates and the schools provided smaller values for the differences between actual and estimated marks. Therefore, from this point onwards, the focus is on the results from the OLS model (b). These are compared with the results obtained using the percentile and z-score methods.

For all units considered and regardless of the different mechanisms used to create the partial absence data (scenarios 1 to 3), the estimated marks calculated via regression imputation provided the best result in terms of the sum of absolute differences between actual and estimated marks.

The percentile and z-score methods provided very similar results. However, it should be noted that the differences between actual and estimated marks were usually smaller when using the z-score method than when using the percentile method (the z-score method yielded smaller total errors in 12 out of 16 tests above). Only for Unit F213 and Component J567/01, and in scenarios 2 and 3, did the percentile method seem to perform slightly better.

Due to the fact that the maximum UMS marks available was not the same for all units (see Table 1), it was not possible to compare the values for the sum of absolute differences across all units in Table 7. However, G322 and J567/01 had the same maximum UMS marks (100 marks) and comparisons were therefore possible. In this case, the measure of discrepancy was smaller in all scenarios and for all methods (particularly in the z-score and percentile methods) in Component J567/01. The marks in this component were estimated using information from Component J567/02, which assesses similar content in the same way: that is, via an external written paper. In contrast, the marks for Unit G322, a written paper, were estimated using information from a coursework unit (G321).

Therefore, when the marks in the units involved in the analysis do not correlate strongly (which is usually the case between coursework and written paper units), neither the z-score method nor the percentile method seems appropriate for estimating the missing marks. The regression imputation method, which includes further information about the candidates, their overall performance and takes into account characteristics of the schools, provides better estimates for the missing marks.

It should be noted that the method itself, rather than the additional information, may be what makes regression imputation a better option for estimating missing marks. However, if we look at the results from OLS model (a), which was based on the same information as the percentile and z-score methods, there were improvements in some cases (particularly F212, F213 and G322) but not in all, and the differences between actual and estimated marks were sometimes much smaller in OLS model (b) than in OLS model (a). This shows that there might be an effect of method but often the effect of the additional data is bigger.

In order to investigate further the effect of using the mark in a coursework unit to estimate the mark in a written paper, missing marks in Unit F212 were estimated leaving out the marks in Unit F213. Table 8 below shows the results.

By comparing the first three rows of Table 7 with the same rows in Table 8, we can see that in the percentile and z-score methods, the differences between actual and estimated marks are smaller (overall) when the coursework marks are ignored. Results for the regression imputation (OLS model [b]) are very similar in both tables.

Therefore, before estimating any missing marks, it seems worthwhile to decide which unit(s)/component(s) of the same specification should be considered and which ones should be discounted. This seems fairly relevant when using the percentile or z-score method, but not so much when using regression imputation.

Table 8: Differences between actual and estimated marks (F212 based on F211)

Unit/ Component	*Missing data generation*	*Sum of absolute differences*		
		Percentile	*z-score*	*Regression imputation OLS (b)*
F212	Scenario 1	842	829	747
	Scenario 2	745	728	666
	Scenario 3	737	730	612

The RMSE statistic was also calculated for all units/components and in all scenarios. The results were consistent with those presented in Table 7, that is, the marks calculated via regression imputation provided the closest estimates (overall) and the percentile and z-score methods provided very similar results.

Correlations between actual marks and marks estimated by the three proposed methods are given in Table 9 below. In this case, the method with the highest correlations would be the best to use.

In all scenarios, correlations were highest when the missing marks were estimated by regression imputation, particularly when including additional data in the models OLS model (b) and quantile regression model (b). The differences between regression imputation and the percentile and z-score methods were bigger when different types of units were involved in the analyses (e.g., Units F212, F213, A267 and, particularly, G322). However, for Unit B601 and Component J567/01 (marks in a written paper were estimated using performance in written papers only), correlations were fairly high, and similarly, independent of the estimation method.

Grading based on estimated marks

In this section, qualification grades based on marks calculated by the three different methods discussed earlier are presented and compared with actual grades.

Firstly, the percentages of candidates who achieved the same estimated grade as the actual grade by each of the three methods considered in this work are displayed in Table 10. This shows that, as was reported for the marks in the previous section, the regression imputation method (either using OLS or quantile regression) provides, overall, the most accurate results.

Table 9: Correlations between actual and estimated marks

Unit/ Component	*Missing data generation*	*Correlations*					
		Percentile	*z-score*	*Regression imputation*			
				OLS		*Quantile regression*	
				(a)	*(b)*	*(a)*	*(b)*
F212	Scenario 1	0.794	0.815	0.864	0.881	0.864	0.880
	Scenario 2	0.627	0.672	0.769	0.840	0.770	0.840
	Scenario 3	0.785	0.778	0.845	0.882	0.845	0.881
F213	Scenario 1	0.516	0.547	0.546	0.569	0.546	0.566
	Scenario 2	0.221	0.199	0.200	0.296	0.200	0.297
	Scenario 3	0.487	0.457	0.457	0.644	0.456	0.650
G322	Scenario 1	0.394	0.412	0.412	0.526	0.412	0.528
	Scenario 2	0.351	0.366	0.366	0.316	0.366	0.316
	Scenario 3	-	-	-	-	-	-
A267	Scenario 1	0.799	0.806	0.808	0.821	0.806	0.818
	Scenario 2	0.734	0.724	0.720	0.751	0.722	0.752
	Scenario 3	-	-	-	-	-	-
B601	Scenario 1	0.837	0.840	0.850	0.855	0.848	0.854
	Scenario 2	0.836	0.838	0.844	0.860	0.843	0.860
	Scenario 3	0.716	0.722	0.714	0.740	0.717	0.744
J567/01	Scenario 1	0.932	0.932	0.932	0.931	0.932	0.931
	Scenario 2	0.924	0.921	0.921	0.920	0.921	0.921
	Scenario 3	0.886	0.886	0.886	0.894	0.886	0.893

Table 10: Percentage of candidates whose estimated overall grade was the same as the actual grade

Unit/ Component	*Missing data generation*	*% achieving the same grade (Number of candidates = 60)*					
		Percentile	*z-score*	*Regression imputation*			
				OLS		*Quantile regression*	
				(a)	*(b)*	*(a)*	*(b)*
F212	Scenario 1	61.7	61.7	65.0	65.5	70.0	62.1
	Scenario 2	50.9	49.1	50.9	52.9	49.1	51.0
	Scenario 3	63.3	65.0	61.7	66.0	65.0	66.0
F213	Scenario 1	83.3	81.7	83.3	81.5	85.0	81.5
	Scenario 2	86.7	83.3	83.3	88.5	85.0	86.5
	Scenario 3	73.3	71.7	80.0	79.1	78.3	79.1
G322	Scenario 1	47.5	44.1	52.5	50.0	50.8	50.0
	Scenario 2	50.0	50.0	61.7	60.3	63.3	62.1
	Scenario 3	-	-	-	-	-	-
A267	Scenario 1	70.0	70.0	70.0	69.5	71.7	69.5
	Scenario 2	65.0	71.7	70.0	69.5	70.0	69.5
	Scenario 3	-	-	-	-	-	-
B601	Scenario 1	68.3	70.0	78.3	76.7	76.7	71.7
	Scenario 2	79.7	83.1	84.7	84.2	84.7	80.7
	Scenario 3	68.1	74.5	74.5	72.3	74.5	72.3
J567/01	Scenario 1	76.3	81.4	79.7	78.2	79.7	78.2
	Scenario 2	76.7	78.3	76.7	80.4	76.7	80.4
	Scenario 3	81.7	80.0	81.7	81.7	81.7	81.7

The percentages of candidates whose estimated overall grades were the same as the actual grades were very similar when using both the percentile and the z-score method.

There were, however, a couple of cases when the regression imputation provided the worst results and either the z-score or the percentile method was the best method in terms of preserving the actual grades. However, given the small numbers of cases used in each analysis, the differences are unlikely to be statistically significant.

Overall grade distributions for the qualifications including the units in Table 10 were also computed and compared with the actual grade distributions. Table 11 shows the average absolute differences of the cumulative percentages between the actual and estimated grades.

Across half of the units and scenarios the regression imputation method (only results for models estimated using OLS regression and including additional data on students and schools are presented in the table) provided the (slightly) highest average, despite being the method providing the best results in terms of the percentage of candidates whose estimated overall grade was the same as the actual grade. However, because only a very small number of candidates have partial absences, the differences in the grade distributions should not be significant in practice.

The absolute differences varied slightly by grade. However this did not appear to be associated with a particular estimation method. For example, the absolute differences were not always bigger at grade A when marks were estimated by regression imputation. This is at odds with the work by Cheung (2009), who reported that the regression method had poor performance in estimating grades at both ends of the distribution.

Table 11: Average absolute differences of cumulative percentages between actual and estimated grades

Unit/ Component	*Missing data generation*	*Average absolute differences*		
		Percentile	*z-score*	*Regression imputation OLS (b)*
F212	Scenario 1	0.003	0.003	0.004
	Scenario 2	0.006	0.005	0.005
	Scenario 3	0.004	0.003	0.004
F213	Scenario 1	0.016	0.016	0.018
	Scenario 2	0.023	0.023	0.024
	Scenario 3	0.016	0.016	0.017
G322	Scenario 1	0.012	0.012	0.024
	Scenario 2	0.017	0.017	0.021
	Scenario 3	-	-	-
A267	Scenario 1	0.344	0.344	0.344
	Scenario 2	0.175	0.227	0.227
	Scenario 3	-	-	-
B601	Scenario 1	0.009	0.008	0.007
	Scenario 2	0.004	0.004	0.004
	Scenario 3	0.005	0.005	0.005
J567/01	Scenario 1	0.002	0.000	0.001
	Scenario 2	0.002	0.003	0.001
	Scenario 3	0.003	0.002	0.003

Conclusions

This work explored the numbers of unit/component marks in GCSE and AS/A level qualifications that were estimated by the OCR awarding body in the June 2015 session and compared three different methods (current and new) for estimating missing marks.

Very small numbers of marks were estimated. In particular, Ofqual found that 99.9% of A levels were graded by examiners in June 2014 (Espinoza, 2015). This was supported by the figures presented in this work, which highlighted that below 0.1% of the AS/A level scripts marked by OCR had estimated marks in June 2015.

Regarding the three methods compared in this research (z-score, percentile and regression imputation), regression imputation seemed to be the most accurate for estimating marks for all units/components and regardless of the different mechanisms used to create the partial absence data. When calculating grades based on estimated marks, the z-score and/or the percentile method were better in terms of preserving actual grades in a couple of instances. However, differences between methods were, in general, small.

In particular, the analyses presented here showed that:

- For the regression imputation method, the two different regression techniques considered (OLS and quantile) provided very similar results.
- For each of the regression techniques above, two different models were estimated. The first model only used the marks in the other units/components that counted towards the qualification; the second model included additional information about the candidates. The models with the additional data provided the best results.
- Although there was an effect of the estimation method on the accuracy of the estimated marks, the results of the analyses carried out here showed that the effect of the additional data was often bigger.
- When the marks in the units/components involved in the analysis do not correlate strongly (which is usually the case between coursework and written paper units/components), neither the z-score method nor the percentile method seems appropriate to estimate the missing marks. Regression imputation provides better estimates.
- Before estimating any missing marks, it seems worthwhile to identify which unit/component or combination of units/components should be considered and which ones should be discarded. Although this is a recommendation for all methods, it seems most relevant when using the z-score or the percentile methods.
- The percentages of candidates whose estimated overall grades were the same as the actual grades were very similar when using both the percentile and the z-score method. The regression imputation method provided, overall, the most accurate results.
- Work by Cheung (2009) found that the z-score method was better than regression (missing mark estimated based on marks in other units only) when comparing unit grades using an average of the absolute differences of cumulative percentages between actual and estimated grades. In a few instances, we found the same results for the overall qualification grade. However, because only a very small number of candidates have partial absences, the differences in the grade distributions would not be significant in practice.

Although regression imputation seems to have several advantages over the other two methods used currently by the UK awarding bodies offering GCSEs and AS/A levels, there is an important limitation to consider. If data is not available for a candidate in one or more of the variables included in the regression models (e.g., overall performance[5], average school performance), an estimated mark is not calculated. There are a couple of solutions in this instance. Firstly, the missing information can be estimated using statistical methods to handle missing data and, once available, the regression imputation proceeds as described in this article. An alternative is to use another method (e.g., z-score or percentile method) in those instances, or use in the imputation only those variables for which there is information.

[5] Although, if a candidate has assessment scores from at least one other component then the overall performance measure should never be missing.

GCSEs and AS/A levels are currently being reformed, with many of the new reformed qualifications available for certification from June 2017. One of the main changes being introduced is the return to linear assessments. As a result, the JCQ has been recently working towards a common approach for how to calculate estimated marks in the new linear qualifications. Alternative methods such as the ones looked at in this research (e.g., z-score, percentile and regression imputation) have been considered in a variety of different research projects carried out by the different UK awarding bodies. The outcomes from such research did not show an outstanding method, but rather very small differences between them (this research shows just a marginal preference for regression imputation, with the performance of the z-score and percentile methods very similar). As the majority of the UK awarding bodies already use the z-score method for unitised specifications, it was agreed by the JCQ that it should be used for the new linear specifications from 2017 onwards.

References

Cheung, C.P. (2009). *Investigating different methodologies for calculating missing marks – examples using data from GCE AS new specifications (Economics and French)*. Internal Report. Cambridge: Oxford, Cambridge and RSA.

Espinoza, J. (2015). A-level results: exam boards 'guesstimating' students' grades. *The Telegraph*. (2015, August 01). Retrieved from: http://www.telegraph.co.uk/education/secondaryeducation/11777405/A-level-results-exam-boards-guesstimating-students-grades.html

JCQ (2016). *Access Arrangements and Reasonable Adjustments 2016–2017*. London: Joint Council for Qualifications. Available online at: https://examining.jcq.org.uk/exams-office/access-arrangements-and-special-consideration/regulations-and-guidance/access-arrangements-and-reasonable-adjustments-2016-2017

Linning, S. (2015). British exam boards forced to 'estimate' hundreds of A-level results each year as students' papers are 'lost in the post'. *Mail Online*. (2015, August 01). Retrieved from: http://www.dailymail.co.uk/news/article-3182362/British-exam-boards-forced-estimate-hundreds-level-results-year-students-papers-lost-post.html

On the reliability of applying educational taxonomies

Victoria Coleman Research Division

Introduction

Educational taxonomies are classification schemes which organise thinking skills according to their level of complexity. They provide a unifying framework alongside common terminology that can be used by educationalists. Primarily, most educational taxonomies focus upon thinking skills that fall within the cognitive domain, although some have also included other domains. Educational taxonomies can have a variety of different applications (Marzano, 2001). First, they can be used to analyse existing educational materials such as learning objectives, curriculum plans, lessons and assessments to ascertain which levels of thinking skills they encompass. Secondly, it is possible to use them as a framework when designing educational materials to ensure that the desired cognitive levels are targeted. They can also be adapted to form an assessment tool themselves, for example, by asking markers whether students have exhibited the required level of thinking during assessment activities. Finally, they can be used to ascertain whether corresponding curriculum objectives and assessment materials align, or whether there is a mismatch in the thinking levels that they are targeting. This can be done both in the context of designing new educational materials and in analysing pre-existing ones. Educational taxonomies can be applied in this way to a broad variety of educational contexts, being adapted according to the specific topic under investigation. This literature review will outline research investigating educational taxonomies and their use in terms of reliability.

Reliability

When discussing educational taxonomies and their application, it is important to consider reliability, as the value of different educational taxonomies is somewhat impacted by reliability constraints. There are various different types of reliability. In the subsequent literature review both inter- and intra-rater reliability are discussed[1]. The amount of consideration given to the rater reliability of educational taxonomies is highly variable and studies specifically investigating it in this context are sparse. There are three broad categories of techniques for assessing rater reliability: consensus estimates, consistency estimates, and measurement estimates (Stemler & Tsai, 2008). Within these there are a number of statistical methods that can be used in order to calculate reliability, and the technique which is selected depends on a number of factors such as the type of reliability being assessed and the nature of the data (McHugh, 2012; Stemler & Tsai, 2008). For example, a greater level of inherent dissimilarity between the items being categorised within a taxonomy will tend to lead to higher values of correlation-based reliability coefficients. However, if nearly all items being assessed fall into one or two categories, then there is little distinction between items and so less chance for raters to display a high correlation between their judgements. Conversely, if nearly all items are within a single category, simple measures (e.g., the percentage of times raters agree with one another) will appear high, as even random placement will result in a high level of agreement.

Within educational taxonomy research there has been a great deal of variation in the statistical measures used and in how the resulting reliability statistics have been interpreted. It must be noted that many of the research articles reviewed did not provide full details of the method used to calculate reliability, which limits our interpretation to some extent. Table 1 summarises the methods used by the studies in this review to calculate reliability.

[1] Rater reliability is frequently referred to using different terms in the literature including coder, assessor and judge in place of rater, and terms such as consistency and agreement in place of reliability. However, for the purposes of this review, the term inter-rater reliability will be used.