Appendix 1 – *continued*

Strategy	*Usage and description*	*Complexity of judgement processes entailed**
Evaluating	When an examiner attends to either all or part of the answer space and must process the content semantically, considering the candidate's response for structure, clarity, and logic or other features the mark scheme deems creditworthy.	Complex
Scrutinising	Only when a candidate's response is unanticipated or wrong. The examiner endeavours to spot the route of the error, and whether a valid substitute to the mark scheme solution has been given. During the process, the examiner considers various aspects of the candidate's answer with the intention of recreating what the candidate was attempting. The examiner may have to deal with a lot of uncertainty and re-read the response several times.	Complex
No response	When there is nothing in the answer space, the examiner checks the answer space a couple of times to confirm there is no answer and then awards 0 marks.	Simple

*Note: when interpreted within the context of dual-processing theories of judgement, 'simple' strategies entail *System 1* (intuitive) judgements, whereas 'complex' strategies entail *System 2* (reflective) judgements.

ASSESSMENT JUDGEMENTS

Capturing expert judgement in grading: an examiner's perspective

Peter King, Cambridge Examiner, **Dr Nadežda Novaković and Dr Irenka Suto** Research Division

Introduction to the study

There exist several methods of capturing expert judgement which have been used, or could potentially be used, in the process of determining grade boundaries for examinations. In a recent study conducted within Cambridge Assessment's Research Division, we sought to explore the judgements entailed in three such methods: (i) rank ordering, (ii) traditional awarding, and (iii) Thurstone pairs. Rank ordering requires judges to make relative holistic judgements about each of a series of up to ten scripts, in order to place them in order of overall quality (Black and Bramley, 2008, Gill *et al*., 2007). Traditional awarding, which is England's current principal grading method (QCA, 2008), utilises limen referencing (Christie and Forrest, 1981; French *et al*., 1988; Greatorex, 2003). Recommendations for grade boundaries are made by a committee of senior examiners based upon absolute judgements of whether selected scripts are worthy or unworthy of particular grades. Finally, like rank ordering, the Thurstone pairs method (Thurstone, 1927a, b) requires judges to make relative holistic judgements about scripts. However, judgements are comparisons of pairs of scripts, rather than rankings of larger series of scripts.

The study was conducted in the context of two contrasting examinations from AS level biology and GCSE English. A key aim was to identify the features of candidates' scripts that affect the judgements made in each of the three methods. To achieve this, sixty experienced examiners were invited to participate in the study (thirty for each subject). Each examiner made judgements about overall script quality, using each method on a different batch of scripts. Additionally, each examiner completed a research task in which he or she was asked to rate a fourth batch of scripts for a series of features, using rating scales devised by the researchers. Subsequent data analysis entailed relating the judgemental data on script quality to the script feature data.

Obtaining an examiner's perspective

Immediately after taking part in the study, one examiner recorded and offered the Research Division his views and experiences of participation. His perspective is the focus of this article. While researchers have many opportunities to report their views, the first-hand experiences of research participants generally receive much less attention, yet perspectives of this nature can be immensely valuable. On some occasions, they can be used to triangulate research findings or provide greater depth and explanation of phenomena. At other times they may prove valuable in informing the design and direction of future research. Furthermore, recruitment of these crucial volunteers and their colleagues for further studies may depend upon research being perceived as meaningful and valid, and affecting policy and practice positively.

The examiner is one of Cambridge Assessment's most experienced examiners. He became an English teacher in 1957 and was appointed a Cambridge examiner for O-levels two years later. Over the past fifty years, he has also been involved in GCSE marking, the moderation of coursework, and the training of examiners, amongst other assessment activities. He has retired as Head of English at a comprehensive school in England, and wrote the following account of his participation as a judge in the study.

A first-hand account of participation

"Just as we are currently asking searching questions about our public examination system, so questions are now being asked about the best methods of assessing candidates' work. This may stem from a variety of reasons: the need to make assessment as economically viable as possible; awareness through research projects that there are valid alternatives to traditional marking and awarding; technological changes that make a reality of reliable on-screen assessment.

These are thoughts that were inspired by my recent involvement in one such project by the Research Division of Cambridge Assessment. The work was carried out entirely at home rather than at an award meeting in Cambridge. It involved four batches each of about twenty scripts from OCR English GCSE Unit 1900, Paper 2431/2 (Non-fiction, Media and Information) for the 2006 and 2007 summer examinations. Each batch required a different approach:

- Rank ordering of Batch 1 scripts.
- Traditional awarding exercise for Batch 2.
- Thurstone pairs (paired comparisons) for Batch 3.
- Rating scripts for individual features for Batch 4.

Such an all-embracing exercise proved thought-provoking, leading me to ask some searching questions after years of traditional assessing of English examination scripts. As someone whose experience included moderation of folders of coursework, where rank order is sacrosanct, Batch 1 posed few problems of placing the scripts in what I considered the correct descending order after reading but not re-marking. Whilst it brought home again the importance of comparison and discrimination between scripts, it seemed to have little advantage over the traditional assessment required for Batch 2 where it is essential to assess each script in relation to specified criteria, with clear descriptors for each band or level in which they are to be placed. The latter approach, however, is extremely time-consuming, requiring an initial close scrutiny of the mark scheme before one feels that one has a complete grasp of its complexity. It is also an approach where the ability to make concise, apt comments (based on the criteria) at the end of each task is at a premium. It should, however, be a highly reliable method of assessment, provided examiners put in this groundwork and don't try to work too quickly – something not easy to guarantee, especially where such work is done in the evening or at the weekend after a highly demanding day or week as a full-time teacher or lecturer. It is a distinct advantage to be retired!

The Batch 2 traditional approach highlighted another possible problem with the holistic approach required of Batch 1 (where the script is not re-marked but considered in its entirety). Holistic approaches still require complete familiarity with a complex mark scheme, something not easily acquired for Batch 2 assessment, before one can have complete confidence in one's judgement.

Thurstone Pairs was a new and attractive approach for me but poses the same problems as suggested for the first holistic exercise. Where it was of particular value was that it involved comparisons between scripts from 2006/2007, with valuable cross-checking of whether standards are comparable year on year. The scripts were cunningly paired, often involving reading the script a second time and comparing it with another new script. I suspect it has more advantages with extended writing papers/exercises than with my paper where different tasks require different approaches and criteria. I can see how it could act as a quick, valuable cross-check of standards where scripts have first been traditionally assessed.

Unlike the three methods requiring judgements about overall script quality, the research task of rating a fourth batch of scripts for a series of features (from mechanical aspects such as spelling or handwriting to questions of relevance, the length of the response or the degree of sophistication or understanding or coherence in the writing) proved to be a slightly unsatisfactory exercise. It was generally not as demanding, failing to involve one fully and leaving one wondering whether one had really done justice to the script by such a fragmentary approach. It was a salutary reminder of how assessors often fail to do justice to a piece of work when they focus on particular features rather than the overall quality.

I concluded it is good to be made to think in different ways about methods of assessment. However, in terms of justice to each candidate, I feel that there are no short cuts in English, and that of the three methods of judging overall script quality, the traditional approach is the fairest. Where such research and re-thinking could be an advantage, however, is if it brought home to hard-pressed English teachers that they need to use a variety of approaches when assessing day to day work (holistic, paired, traditional) rather than predominantly focussing on detailed 'correcting' of pupils' work."

Findings

The data analysis for this project has been complex and lengthy. It is intended that the findings will be disseminated in a subsequent report.

References

Black, B. & Bramley, T. (2008). Investigating a judgemental rank-ordering method for maintaining standards in UK examinations. *Research Papers in Education*, **23**, 3, 357–373.

Christie, T. & Forrest, G.M. (1981). *Defining public examination standards*. Schools Council Research Studies. London: Macmillan Education.

French, S., Slater, J. B., Vassiloglou, M. & Willmott, A. S. (1988). *The role of Descriptive and Normative Techniques in Examination Assessment*. In: H. D. Black and B. Dockrell (Eds.), *Monograph of Evaluation and Assessment Series No. 3*. Edinburgh: Scottish Academic Press.

Gill, T., Bramley, T. & Black, B. (2007). *An investigation of standard maintaining in GCSE English using a rank-ordering method*. Paper presented at the British Educational Research Association Conference, London, September 2007.

Greatorex, J. (2003). *What happened to limen referencing? An exploration of how the Awarding of public examinations has been and might be conceptualised*. Paper presented at the British Educational Research Association Conference, Edinburgh, September 2003.

QCA (2008). *GCSE, GCE, VCE, GNVQ and AEA Code of Practice2007/8*. London: Qualifications and Curriculum Authority.

Thurstone, L.L. (1927a). Psychophysical analysis. *American Journal of Psychology*, **38**, 368–389. Chapter 2 In: L. L. Thurstone (1959), *The measurement of values*. Chicago: University of Chicago Press.

Thurstone, L.L. (1927b). A law of comparative judgment. *Psychological Review*, **3**, 273–286.

Investigation into whether z-scores are more reliable at estimating missing marks than the current method

Peter Bird Operational Research Team, OCR

Context

The awarding bodies in the UK use similar, but slightly different, methodologies for assessing missing marks (i.e. marks for candidates who are absent with good reason). In an attempt to standardise the process across awarding bodies, a z-score method of estimating missing marks (as used by other awarding bodies) was investigated to see if it was better than the current proportional estimation method being used by OCR. The proportional method requires the available marks for a candidate to be used in calculating the missing mark, and therefore this method is straightforward to apply operationally. Any new method would also be constrained by what can be achieved operationally at what is already a very busy time of year for processing. The aim of this article is to compare the two methods for a sample of specifications, to highlight any issues and differences in the accuracy of estimating marks. Further, more in depth work, could then be undertaken if required.

Introduction

Two GCSE specifications and three A-level specifications were used to investigate whether a z-score estimation method was better than the current estimation method in use. The subjects were chosen because of their very different characteristics. This study was designed to be an exploration of the likely issues and problems from each method before a more in-depth analysis was carried out.

The 'current proportional method', which has been in use for many years, assumes that a candidate will perform equally well on the unit/components that they are missing as they did on the units/components for which they have marks. The 'z-score method' assumes that the relative position of a candidate's mark in relation to all other candidates taking the same unit/component stays the same for both unit/components. In short: i) the existing method assumes the same proportional score in relation to the maximum mark on the missing component(s) as on the components taken; ii) the z-score method assumes the missing mark lies the same number of standard deviations from the mean on the missing component as on components taken.

Each method was compared by treating in turn all candidates as having missing marks. Estimates were then calculated. This was repeated for each unit/component within each specification.

At OCR, missing marks have been estimated for many years on the assumption that a candidate performs equally well on the unit/components for which we have marks for them, as they do on the missing unit/component. The reliability of this method relies on assuming there is a good correlation between the unit(s) being predicted from and to, and that the distribution characteristics of each unit are similar. This method does not take into account whether the marks already achieved come from a distribution with the same distributional characteristics as the one which is being estimated, that is, the obtained mark may have come from a skewed distribution, such as a coursework unit, or from a tiered paper, and the estimate may be required for a unit which has a bell-shaped distribution.

The new proposed method of using z-scores is a method which takes into account how well the candidate for which we are estimating a missing mark has performed on other components in relation to all other candidates taking the same unit/component. It effectively gives a higher z-score to a candidate who has achieved a mark in the top end of the mark distribution, and similarly a lower z-score to a candidate who achieved a mark at the bottom end of the mark distribution. For a normal distribution we would expect 68% of candidates to lie within the mean +/- one standard deviation, and 95% of candidates to lie within the mean +/- two standard deviations. A mark is transformed to a z-score by subtracting the mean and dividing by the standard deviation from the distribution it comes from.

Example of applying both methods

Specification with three components. Candidate has component 3 missing.

Component	*Mark Achieved*	*Max*	*Mean**	*Std Dev**	*Calculated z-score*
1	12	40	20	10	=(12-20)/10= -0.8
2	17	50	25	12.5	=(17-25)/12.5= -0.64
3	Missing	30	15	7.5	

* assume bell shaped distributions

Current method to predict missing mark on component 3

$$= \frac{\text{Marks gained on components 1 \& 2 } = 12+17}{\text{(Max Mark on component 1 \& 2 } = 40+50} \times \text{(Max mark on missing component 3 } = 30)$$

= (29/90) × 30 = 9.66 = rounded to 10 marks.

Z score method to predict missing mark on component 3

= [(combined z-score of component 1 & 2) × std dev of component 3] + (mean of component 3)

= [(-0.71) × 7.5]+15=9.675 = rounded to 10 marks.

Where combined z-score component 1 & 2

= [z-score component 1 × (max component 1) / (max component 1+2)) +
[z-score component 2 × (max component 2) / (max component 1+2))]

= [(-0.8 × (40/(40+50))) + (-0.64 × (50/(40+50)))] =
(-0.35) + (-0.35) = -0.71

By using bell shaped distributions for all components with the mean set at half the maximum marks and the standard deviation set at half the mean mark, the estimates for both methods came out very similar.

Effect of z-score process

To see the effect of the process, random data have been generated to create an example of a typical written paper mark distribution with mean 50, standard deviation 15. These have then been converted to z-scores (Figure 1 below).

A ceiling at a mark of 60 was introduced to create a skewed distribution as might be seen in coursework mark distributions. This produced a mean of 43.4 and standard deviation of 11 (below).

The effect of using coursework to predict a mark on the written paper in the example above is that even if a candidate achieves the maximum mark on coursework, they are effectively capped for their estimated mark on the written paper to about 70 out of 100 (because the maximum z-score they can achieve is around +1.5).

The effect of using written papers to predict coursework in the example above is that anyone achieving over approximately 70 marks on the written paper will be estimated as achieving the maximum mark on the coursework.

Combining units/components

Using a combination of different types of mark distribution is more likely to produce less reliable mark estimates than estimating using similar types of distribution. In order to combine z-scores from different units/components to create one z-score, the individual z-scores are weighted according to the relative weightings of each unit/component to each other.

For example,

If a candidate's marks produced z-scores of +1 and +1.5 on units with weightings of 20% and 30% respectively, the combined z-score is $[(+1 \times 20)/(20+30)]+[(+1.5 \times 30)/(20+30)]=(+0.4)+(+0.9)=+1.3$.

Issues with cohorts used for estimating

Coursework marks may be used from a distribution which contains both foundation and higher tier candidates so the mean and standard deviations would not be truly representative of a particular tier cohort. In a unitised scheme, you cannot guarantee the cohort from which an estimate is obtained is the same as the original cohort for the missing unit, particularly with early takers or re-sitters being included. For this analysis it was assumed any mark estimates would be based on the distribution of the missing unit within the same session as the aggregation of unit results was requested.

The more the cohort used for prediction varies, the more you would anticipate that the reliability of the estimation will decrease. In the examples shown so far, these did not involve UMS marks. However, when UMS marks are used for estimating other UMS marks we have to bear in mind the marks have already been subjected to some 'stretching and squeezing' across the mark ranges. Comparisons of the differences in z-scores were looked at between those derived from weighted raw marks

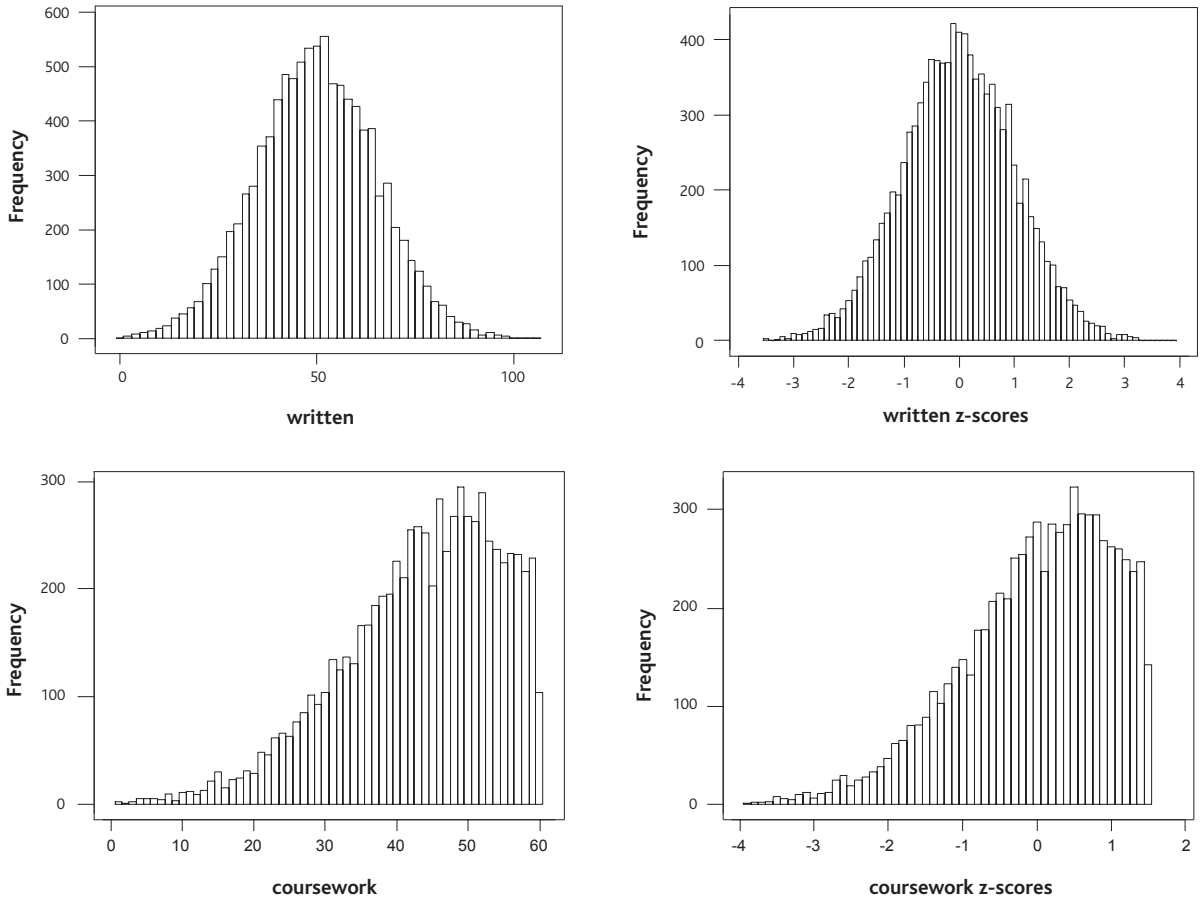


Figure 1: Effect of z-score process

Table 1: Examples of GCSE and A-level differences in cohort/prediction method

Specification type	*Predict z-scores from*	*Map z-scores onto*	*Cohort*
GCSE Linear (non tiered)	Weighted marks from Component(s)	Weighted Mark for missing component	Exactly the same cohort used to predict from and to.
GCSE Linear (tiered)	Written Component(s) from relevant tier and/or coursework	Weighted Mark for missing Component	Exactly the same cohort used to predict from and to. Z-scores could be distorted if coursework tier breakdowns not available.
GCSE Unitised (tiered)	Weighted unit marks for session these were sat in	UMS Mark for unit in aggregation session	Most likely not the same cohort used to predict from and to. Z-scores could be distorted if coursework tier breakdowns not available.
GCSE Unitised (untiered)	Weighted unit marks for session these were sat in	UMS Mark for unit in aggregation session	Most likely not the same cohort used to predict from and to
GCSE Unitised (Linear)	Weighted unit marks for session these were sat in	Weighted/UMS Mark for unit in aggregation session	Exactly the same cohort used to predict from and to. Z-scores could be distorted if coursework tier breakdowns not available.
A/AS-level (missing AS unit)	Weighted AS unit marks for session these were sat in.	UMS Mark for unit in aggregation session	Most likely not the same cohort used to predict from and to
A-level (missing A2 unit)	Weighted A2 unit marks for session these were sat in	UMS Mark for unit in aggregation session	Most likely not the same cohort used to predict from and to.
Missing component within (GCSE or A-Level)	Weighted marks from Component(s)	Weighted Mark for missing component	Exactly the same cohort used to predict from and to unit

and those derived from UMS marks. These showed that for the majority of candidates there are no differences, although the z-scores varied by (+/-) 0.1 to 0.2 for approximately 10–25% of candidates.

For any readers who are unfamiliar with the concept of the Uniform Mark Scale (UMS), an excellent explanation is found in Gray and Shaw (2009).

To improve reliability of estimating for A-level you might want to look only at the best marks from all units of the candidates who are aggregating. Table 1 above outlines differences in the cohort used for different specification types and where reliability issues may exist.

Comparison of different estimation methods

In order to compare different estimation methods, missing marks were created where valid marks already existed for entire units/components, this then allowed comparisons of the estimation accuracy of each method. GCSE linear (tiered), GCSE unitised (untiered) and A/AS-level specifications were used in analysis to look at any differences between tiered/uncapped and UMS conversion specifications. To do this the following assumptions were made:

- Only candidates with complete profiles of marks were included.
- Where options exist within units, the mark used to calculate the z-score is the final weighted mark.
- The z-score is calculated from the unit in the session from which it counted towards aggregation.
- Estimation of unit UMS mark is based on using z-scores from the unit UMS distribution of missing mark in June 2007 (i.e. aggregating session).
- Where optional units exist, the estimation will be based on the marks each candidate has achieved on the units taken.
- Missing AS units are only estimated on AS units.
- Missing A2 units are only estimated on A2 units.
- Very small entry units are excluded.
- Z-score calculations were calculated using data which are shown on our exams processing system.

Estimating marks for candidates aggregating GCSE Geography 1987 in June 2007

GCSE Geography 1987 was used to evaluate the effectiveness of each estimation method as it contains a good mix of distribution types, a large number of candidates and two tiers. Candidates take either Foundation or Higher option and components as below:

Foundation: Component 1 (Foundation) + Component 3 (Foundation) + Component 5 (coursework)

Higher: Component 2 (Higher) + Component 4 (Higher) + Component 5 (coursework)

Summary statistics for each component are shown in Table 2. The foundation option papers are both skewed as candidates tend to get higher than half marks whereas the higher tier candidates' marks are well dispersed on the written paper but skewed on the coursework. The correlation between written papers is higher than between written paper and coursework which makes this a very 'typical' specification. The correlations between component 1 and 3 is +0.71;

Table 2: Summary statistics for the weighted marks for GCSE Geography 1987

Component	*01*	*02*	*03*	*04*	*05 (F=Found, H=Higher)*
MEAN	48.32	49.49	33.09	32	24.5(F)/38.8(H)
STD	12.24	11.65	8.43	7.59	8.9(F)/7.7(H)
N	17271	20591	17271	20591	17271(F)/20591(H)
MAX	90	90	60	60	50
SKEW	-0.50	0.0	-0.59	+0.1	-0.12(F)/-0.69(H)

between 2 and 4 is +0.59; and all remaining correlations are between +0.4 to +0.46.

For each candidate, an estimation of each of their marks was calculated in turn using the other available marks, that is, effectively treating each candidate as having a missing mark. Component 01 was then estimated from marks on components 03 and 05; component 02 was then estimated from marks on components 04 and 05, etc. This was carried out for both the current estimation method and the z-scores estimation method.

COMPONENT 01 (Foundation Written Paper)

The graphs in Graph 1 below show two box plots. The first plots the differences between the estimated written marks on component 1 (using the current estimation method based on component 3 [written paper] and component 5 [coursework]) and the actual marks the candidates achieved. The second shows the same estimation, but using a z-score methodology instead. The vertical axis shows the differences (estimated-actual) and the horizontal axis shows the actual mark achieved. A positive difference shows where the estimation process was over-estimating the mark and a negative difference where it was under-estimating the mark.

The edges of each box for each mark point show where the 25 and 75 percentiles of candidates' marks lie between and the horizontal line within the box is the 50 percentile point. The lines extend to contain 90% of candidates' marks. A box plot of the differences between estimating marks using the z-score method and the actual marks is shown in Graph 2. Both Graph 1 and Graph 2 are very similar, thus both methods produce very similar outcomes although the widths of the 25 and 75 percentiles are marginally smaller using the z-score estimation method.

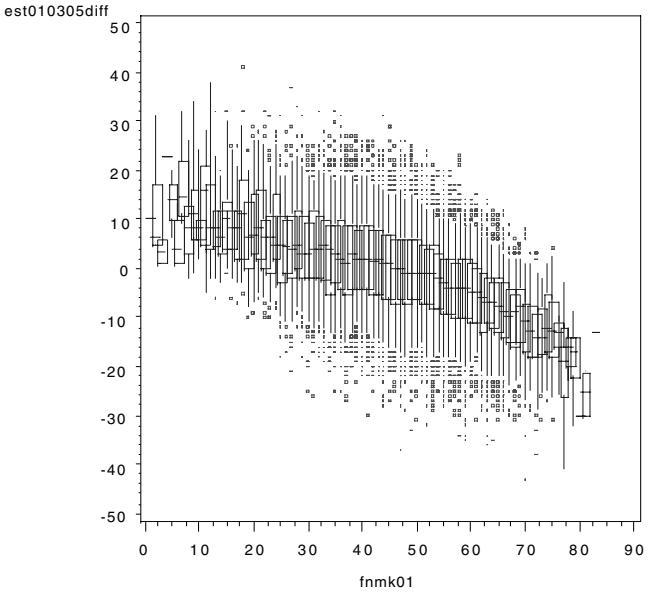
Both box plots show that a candidate would most likely achieve a higher estimated mark than they would have achieved if their actual mark was below the mean mark and a lower estimated mark than they would have achieved if their actual mark was above the mean mark. The differences vary more in magnitude towards the upper end of the mark range.

Using Linear Regression (as a possible method), it is possible to effectively scale/transform the marks in such a manner that the variation on any mark is minimised once all mark estimations have been calculated. A box plot of the differences between estimating marks using the z-score method and then applying a linear regression scaling and the actual marks is shown in Graph 3.

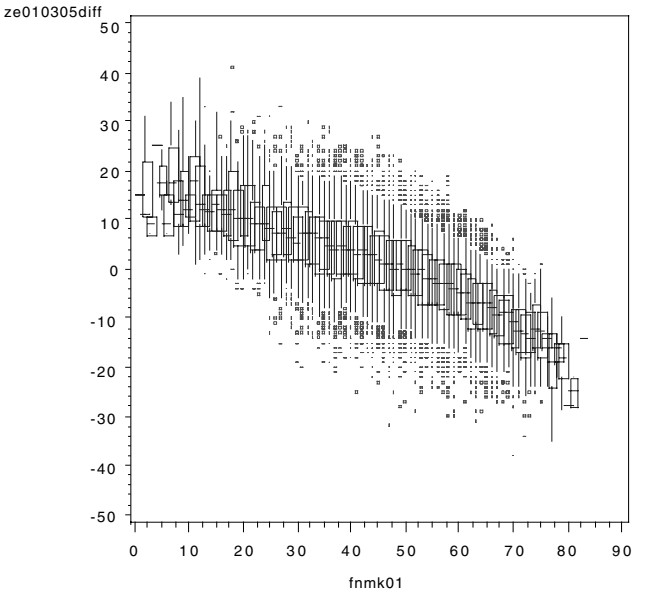
A simple linear regression line was calculated from the differences in Graph 2 treating *ze010305diff* as the dependent variable and *fnmk01* as the independent variable. All differences were then adjusted by subtracting the outcome of this line of best fit for each actual mark.

Using this method, we would be around 90% confident that most mark estimates are within a certain mark range. In this example, the majority of estimates would lie within approximately 10 marks of their actual mark. If this were to be applied to the current estimation method, it would produce similar results.

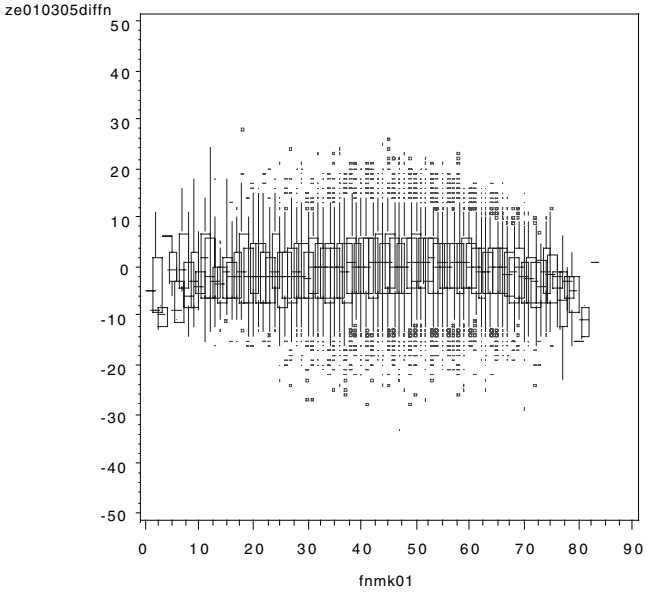
A summary of the differences of the three methods estimation (est), z-scores estimation (ze) and z-scores estimation followed by linear regression scaling (zen) is shown in Table 3. For this 'closed linear cohort' the average z-score estimated difference is 0 marks as by definition transforming to z-scores would do this. Comparisons of the 10, 25, 50, 75 and 90 percentile differences show that as each method is applied, the size of the errors between estimated and actual marks generally decreases.



Graph 1: Plot of the differences between (estimated–actual) against actual mark for component 01 (using current estimation rules)



Graph 2: Plot of the differences between (estimated–actual) against actual mark for component 01 (using Z-scores)



Graph 3: Plot of the differences between (estimated–actual) against actual mark for component 01. (Using linear regression to scale marks after z-score estimation has taken place)

Table 3: Summary of differences between estimated and actual marks for each estimation method for Geography 1987/01

				Percentile				
method	*comp*	*mean*	*std*	*10%*	*25%*	*50%*	*75%*	*90%*
est	01	-1.2	9.9	-14	-8	-1	5	11
ze	01	0.0	9.1	-12	-6	0	6	12
zen	01	0.0	7.5	-10	-5	0	5	10

COMPONENT 02 (Higher Written Paper)

This process was repeated for estimating the marks on component 02 using z-scores from the marks on component 4 (written) and 5 (coursework). A box plot of the differences between estimating marks using the current method and the actual marks is shown in Graph 4 and differences between estimating marks using the z-score method and the actual marks is shown in Graph 5.

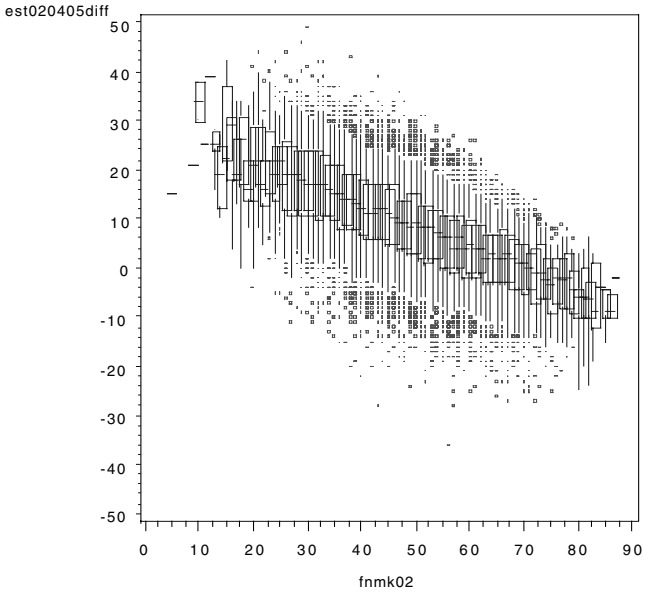
The current estimation process estimates more marks above their actual marks, whereas the z-score method ensures approximately the same number of mark estimates are above and below their actual marks. In contrast to component 1, you do not see the 'dipping' of the plot towards the end of the mark range so the size of the differences are more proportional across the entire mark range.

This example shows how using marks from a coursework distribution (which have a high mean in relation to the maximum mark) as part of the prediction for a written paper (where the mean is closer to half the maximum mark) will over-estimate the marks under the current methodology.

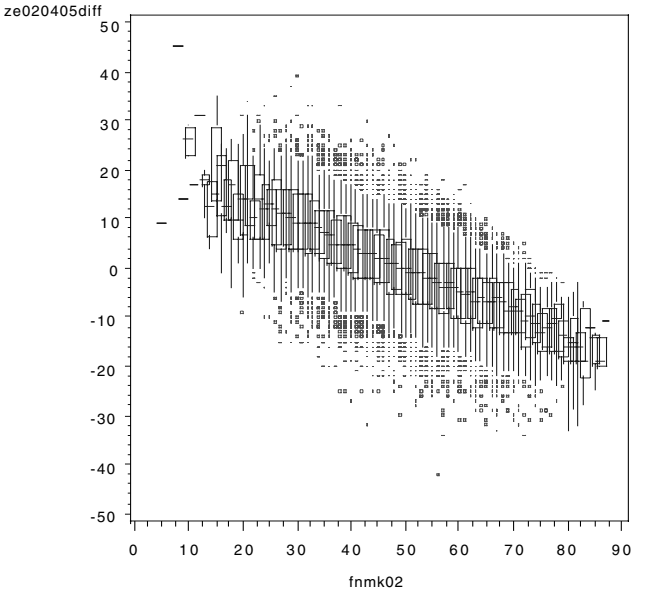
Linear Regression was used to scale/transform the marks in such a manner that expected variation on any mark was minimised once all mark estimations have been calculated. A box plot of the differences between estimating marks using the z-score method (and then applying a linear regression scaling) and the actual marks is shown in Graph 6.

Using this method, we would be reasonably confident that the mark estimate is within approximately 10 marks of their actual mark. If this were to be applied to the current estimation method, it would produce similar results.

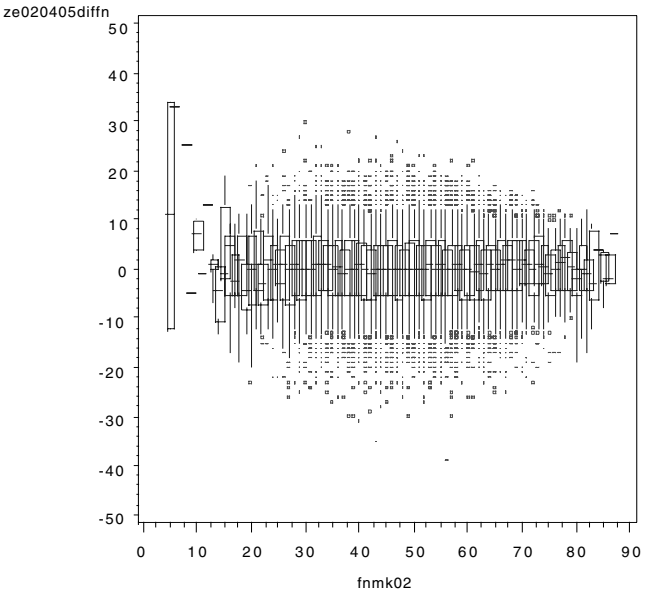
A summary of the differences of the three methods estimation (est), z-scores estimation (ze) and z-scores estimation followed by linear regression scaling (zen) is shown in Table 4. Comparisons of the 10, 25, 50, 75 and 90 percentile differences show that as each method is applied, the size of the errors between estimated and actual marks generally decreases.



Graph 4: Plot of the differences between (estimated–actual) against actual mark for component 02. (Using current estimation rules)



Graph 5: Plot of the differences between (estimated–actual) against actual mark for component 02. (Using Z-scores)



Graph 6: Plot of the differences between (estimated–actual) against actual mark for component 02. (Using linear regression to scale marks after z-score estimation has taken place)

Table 4: Summary of differences between estimated and actual marks for each estimation method for Geography 1987/02

				Percentile				
method	*comp*	*mean*	*std*	*10%*	*25%*	*50%*	*75%*	*90%*
est	02	8.4	9.7	-4	2	8	15	21
ze	02	0.0	9.4	-12	-6	0	6	12
zen	02	0.0	7.6	-10	-5	0	5	10

COMPONENTS 03, 04 and 05

For summary data of the differences seen for the remaining components please contact the author.

In summary, it seems that for 'closed cohort' linear specifications, z-scores would ensure the mean difference between the estimated and actual mark is zero. Any deviations away from this would be balanced positively and negatively. With the current estimation method we cannot guarantee this unless we check using data from all candidates, and make any necessary mark transformations to make this so.

Estimating written papers component 01 and 02 based on written papers 03 and 04 only

It was interesting to try to estimate a written paper mark using only the mark from the other written paper taken so the effect of not using coursework marks for estimation could be seen. Table 5 shows the differences from estimating component 01 from component 03 only, and component 02 from component 04. Only some data are shown here.

This produced slightly better estimates as we might expect. In particular, the mean difference dropped from +8.4 to -1.2 for component 2. In terms of the range of differences seen, component 1 had less large differences at the top end of range and component 02 produced a more even number of positive and negative differences, similar to those seen with the z-scores method.

Table 5: Summary of differences between estimated and actual marks for each estimation method for Geography 1987/01/02 (estimated from written paper only)

				Percentile				
method	*comp*	*mean*	*std*	*10%*	*25%*	*50%*	*75%*	*90%*
est	01	1.6	9.4	-10	-5	2	8	14
ze	01	0.0	9.3	-12	-6	0	6	12
zen	01	0.0	8.6	-11	-6	0	6	11

				Percentile				
method	*comp*	*mean*	*std*	*10%*	*25%*	*50%*	*75%*	*90%*
est	02	-1.2	10.4	-14	-8	-1	6	12
ze	02	0.0	10.5	-13	-7	0	7	14
zen	02	0.0	9.4	-12	-7	0	6	12

A box plot of the differences between estimating marks using the current method and the actual marks for component 01 using component 03 and then using components 03 and 05 are shown in Graphs 7 and 8 respectively below.

Estimation of marks on A-level Physics 7883 for those candidates aggregating in June 2007

Overview

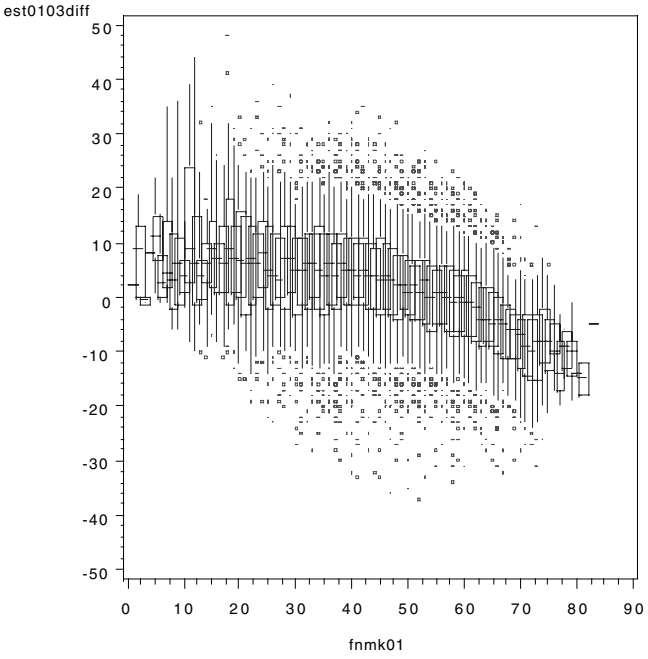
A-level Physics was used to evaluate the effectiveness of each estimation method as it contains reasonable bell shaped distributions, a reasonable number of candidates, and a range of unit types including compulsory/optional and written/coursework or practical. Only A2 units were used for estimation to minimise re-sit effects. 40% of candidates chose to take unit 2824 in both January 2007 and June 2007, whereas less than 5% and 1% did for units 2825 and 2826 respectively. The correlations between units' marks were also all fairly consistent at approximately +0.7 to +0.8. The specification is made up of three AS units 2821–2823 and three A2 units 2824–2826.

Estimation of mark on unit 2824, 'Forces, Fields and Energy'

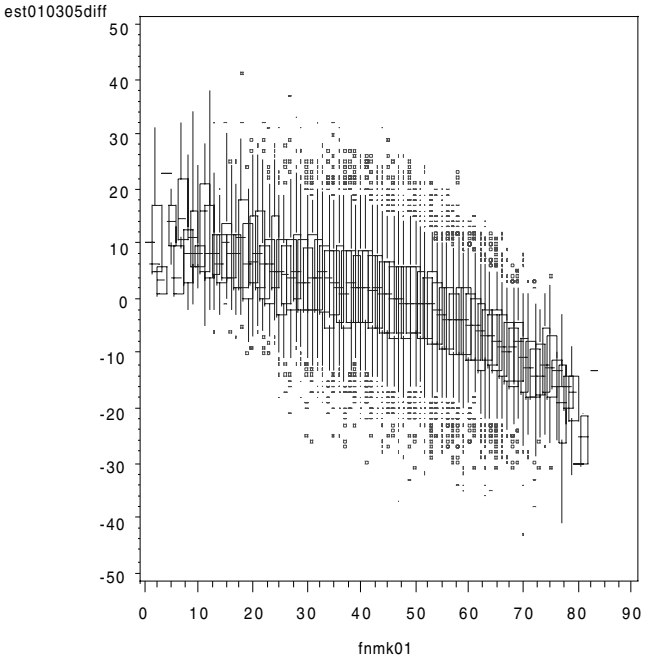
Box plots of the differences between estimating marks using the current method and the actual marks for unit 2824 are shown in Graphs 9 and 10 respectively. This unit is estimated using unit 2825 'Options in Physics' which gives candidates five choices in paper topic and unit 2826 'Unifying concepts' which includes either coursework or practical. In this example, the z-scores method underestimates the actual mark and is prone to slightly more error in estimates across the mark range.

Adjusting for the slope of the differences (zen) makes very little difference to the reliability of the estimated marks calculated using the current estimated method (est); this is shown in Table 6.

Please note that the mean of the differences using the z-scores method is not zero. This seems be an effect of not using a 'closed cohort', that is, z-scores might be pulled from more than one session, where other candidates not aggregating exist and previous attempted marks exist, and these are mapped onto the final aggregation session unit distribution which again may include candidates not aggregating.



Graph 7: Plot of the differences between (estimated–actual) against actual mark for component 01 (Using current estimation rules using component 03 only to estimate from)



Graph 8: Plot of the differences between (estimated–actual) against actual mark for component 01 (Using current estimation rules using components 03 and 05 to estimate from)

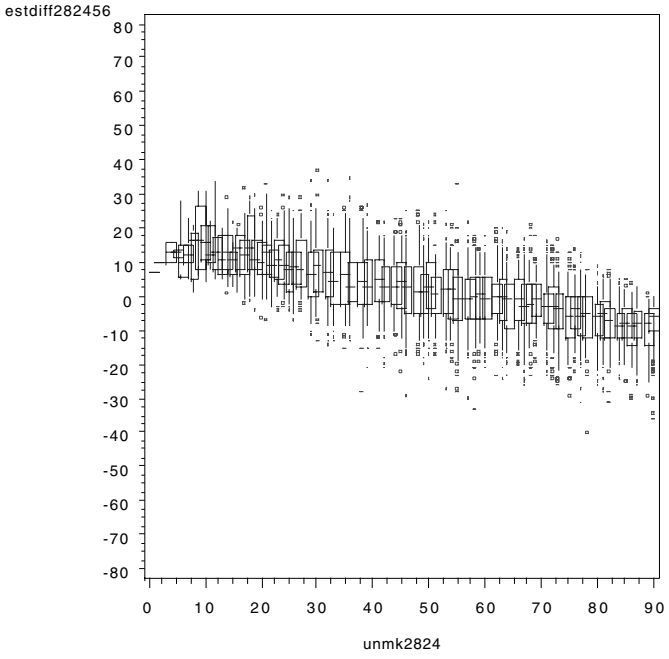
Table 6: Summary of differences between estimated and actual marks for each estimation method for unit 2824

				Percentile				
method	*comp*	*mean*	*std*	*10%*	*25%*	*50%*	*75%*	*90%*
est	2824	0.28	10.13	-12	-7	0	7	13
ze	2824	-4.95	10.48	-18	-12	-5	2	9
zen	2824	-0.01	9.71	-12	-6	0	7	12

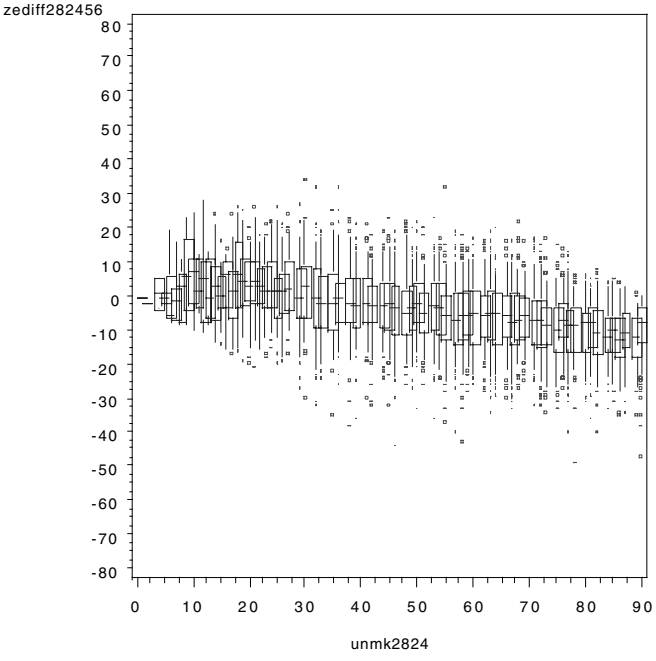
Estimation of mark on unit 2825, Options in Physics

Box plots of the differences between estimating marks using the current estimation method and the actual marks for unit 2825 are shown in Graphs 11 and 12 respectively. Unit 2825 is estimated using unit 2824 and 2826.

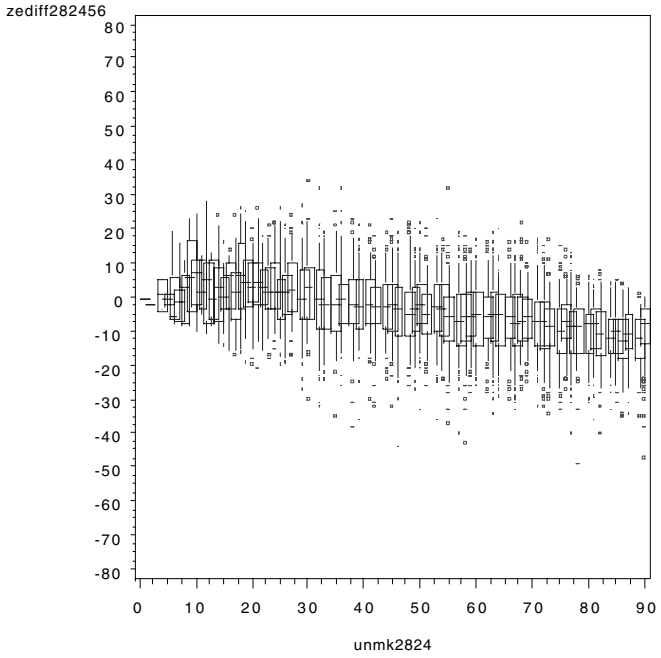
For this unit, the z-scores method is vastly better at estimating the marks than the current estimation method as on average it is only over estimating by 2 rather than 4 marks as shown in Table 7.



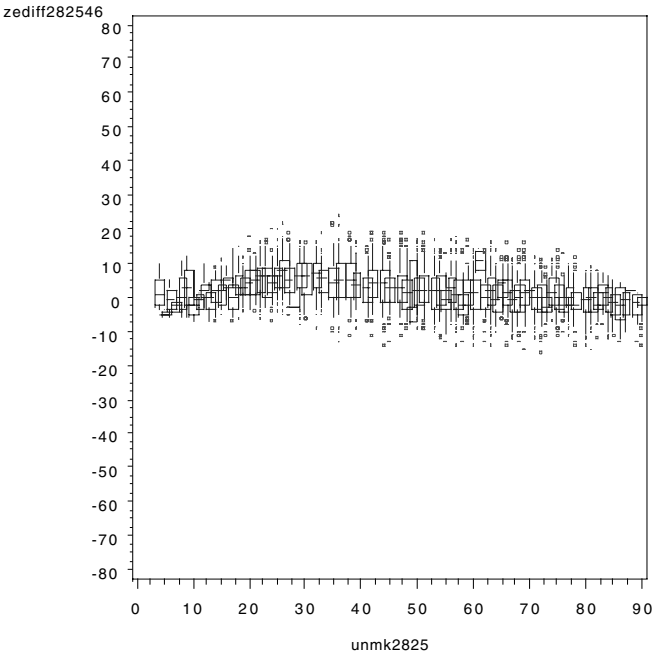
Graph 9: Plot of the differences between (estimated–actual) against actual mark for unit 2824. (Using current estimation rules)



Graph 10: Plot of the differences between (estimated–actual) against actual mark for unit 2824. (Using Z-scores)



Graph 11: Plot of the differences between (estimated–actual) against actual mark for unit 2825 (Using current estimation rules)



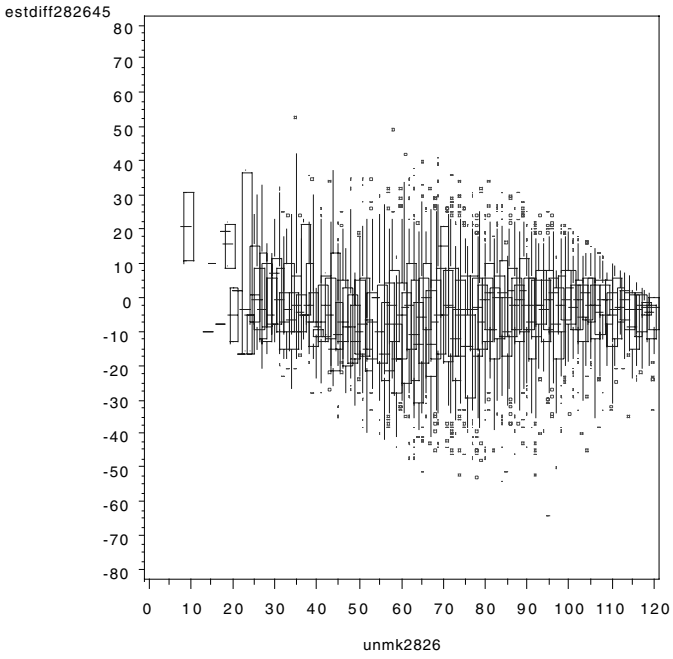
Graph 12: Plot of the differences between (estimated–actual) against actual mark for unit 2825 (Using Z-scores)

Table 7: Summary of differences between estimated and actual marks for each estimation method for unit 2825

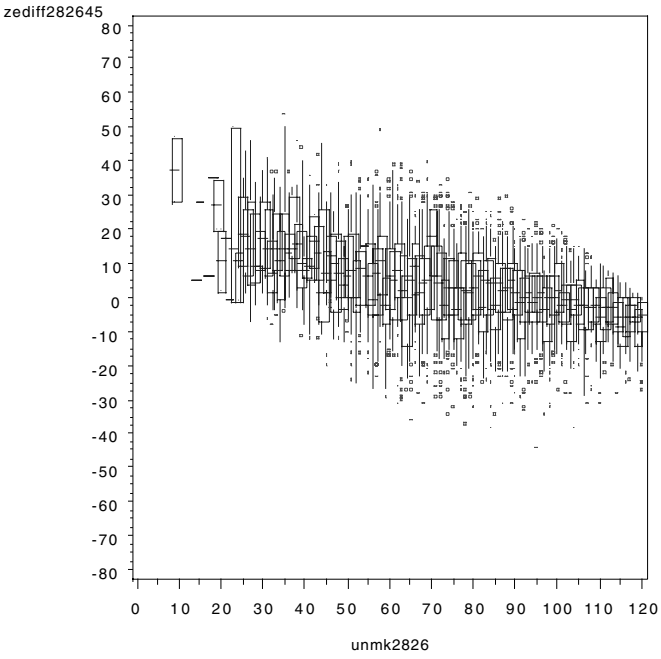
				Percentile				
method	*comp*	*mean*	*std*	*10%*	*25%*	*50%*	*75%*	*90%*
est	2825	3.19	10.56	-10	-4	3	10	17
ze	2825	1.92	5.74	-5	-2	1	6	10
zen	2825	-0.02	5.40	-7	-4	0	4	7

Estimation of mark on unit 2826 'Unifying concepts in Physics'

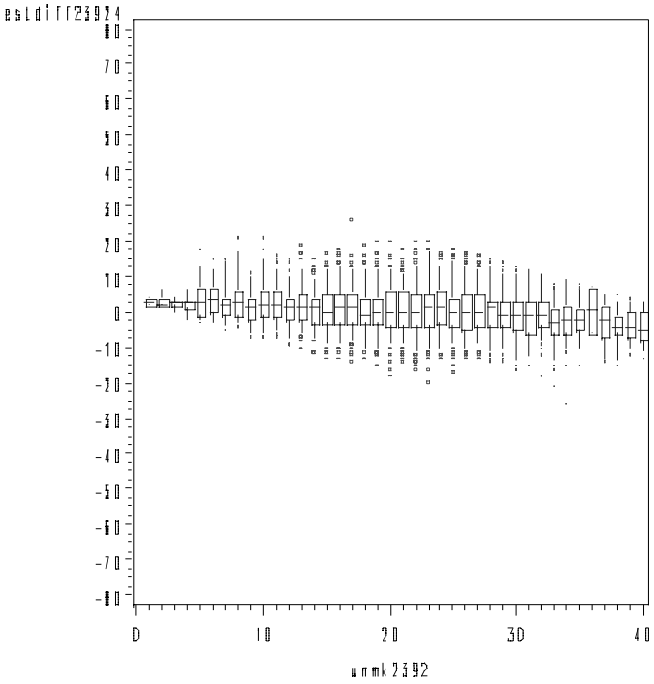
Box plots of the differences between estimating marks using the current method and the actual marks for unit 2825 are shown in Graphs 13 and 14 respectively. Unit 2826 is estimated using units 2824 and 2825. For unit 2826, the estimation of the marks using both methods varies more considerably than the previous units across the entire mark range, as we might expect, as this contains some centre assessed work (Table 8).



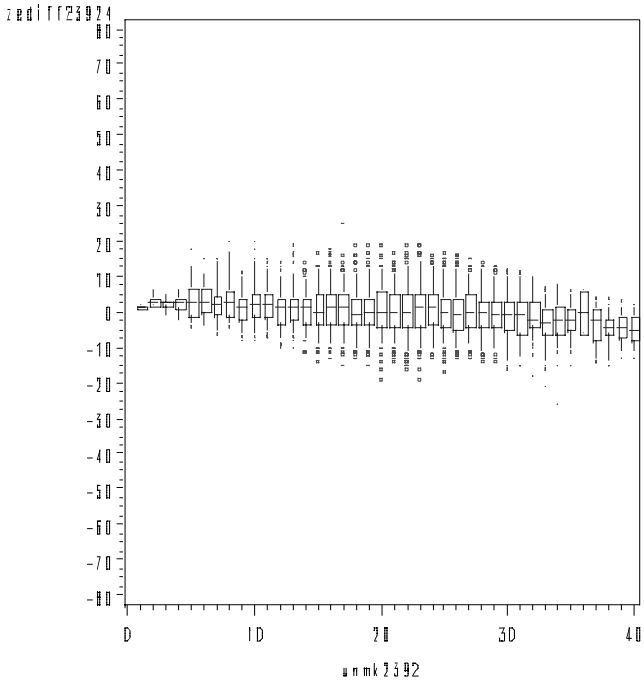
Graph 13: Plot of the differences between (estimated–actual) against actual mark for unit 2826 (Using current estimation rules)



Graph 14: Plot of the differences between (estimated–actual) against actual mark for unit 2826 (Using Z-scores)



Graph 15: Plot of the differences between (estimated–actual) against actual mark for unit 2392 (Using current estimation rules)



Graph 16: Plot of the differences between (estimated–actual) against actual mark for unit 2392 (Using Z-scores)

The z-scores method tends to be more reliable at estimating at the top end of mark range but over-estimates at the bottom end. For this unit, the z-scores method looks more reliable.

Table 8: Summary of differences between estimated and actual marks for each estimation method for unit 2826

				Percentile				
method	*comp*	*mean*	*std*	*10%*	*25%*	*50%*	*75%*	*90%*
est	2826	-4.07	14.70	-24	-14	-3	6	14
ze	2826	2.72	12.35	-13	-5	2	11	18
zen	2826	-0.01	11.47	-15	-7	0	8	14

From the Physics units, it seems that estimating coursework from written papers is always going to be more prone to error as some candidates prefer written papers to coursework. It also seems that estimating based on an optional paper (unit 2825) produces slightly less accurate estimates than those based on a compulsory paper (unit 2824).

A compulsory paper is a measure of the candidates' abilities compared to each other whereas the relative positioning of candidates for a paper which has been chosen may allow more variation, particularly in the smaller entry option papers. The re-sitting of unit 2824 most likely allows any candidates who were not in their correct relative position the first time they sat the unit to improve their z-score for this unit.

A-level French 7861 and AS Business Studies 3811

Units in A-level French 7861 and AS Business Studies 3811 were also estimated. Please contact the author for summary statistics. In French, the estimates were mixed as each unit is testing very different traits, speaking, listening, reading and writing. For some units the current estimation method was better, in others the z-scores method was better, but it seems estimating from distributions with high mean marks in relation to the max mark to distributions with mean marks closer to the half the max marks tends to over-estimate, and vice-versa. In Business Studies the estimates for both methods were very similar for all AS units.

GCSE Religious Studies 1030 (Short course)

Overview

GCSE Religious Studies (short course) contains ten papers/units 2391–2400. Each candidate must take two of these units. There are no tiers, no capping, and each paper produces similar looking distributions with correlations between marks of +0.7 and +0.8 for those candidates aggregating. Estimation of units will therefore be dependent on paper choice.

Unit 2392 – 'Christian Perspectives'

Box plots of the differences between estimating marks using the current method and the actual marks for unit 2392 are shown in Graphs 15 and 16 respectively, in these graphs the marks are estimated using unit 2394. The graphs show that the estimation of the marks using the current estimation method is very close to those estimated by the z-score method. The summary statistics for estimating all marks on this unit using the corresponding unit which was taken are shown in Table 9. In this unit, both the current and z-score methods on average underestimated the marks by around 2 marks. Very similar summaries of differences are found on all units in this specification.

Table 9: Summary of differences between estimated and actual marks for each estimation method for unit 2392 (estimated from other available unit)

				Percentile				
method	*comp*	*mean*	*std*	*10%*	*25%*	*50%*	*75%*	*90%*
est	2392	-2.27	5.67	-9	-6	-2	1	5
ze	2392	-2.25	5.64	-9	-6	-2	2	5
zen	2392	0.00	5.50	-7	-4	0	4	7

Conclusion

There is no perfect system when it comes to estimating marks, as candidates perform differently on different units/components. The current estimation process and the z-scores method both rely on the correlation between units/components being as close to one another as possible, but in practice this is never met. The z-scores method does take into account the relative positioning of candidates in respect to other candidates but it is also affected by different shaped distributions and estimates can be artificially capped. It does, however, try to address the over-inflating of written paper marks where a skewed coursework distribution is used to estimate these.

On linear specifications, z-scores would ensure the mean difference between the estimated and actual mark is zero and thus the direction of any errors in estimating marks would be balanced both positively and negatively across the mark range. This cannot be guaranteed with the current estimation method. However, for unitised schemes (which are continuing to increase in number) it is less clear, as in some cases the estimates were very similar; in some cases better and in some cases worse. This is very much dependant on the types of units, correlations between units marks and distribution types.

Unitised schemes by their nature allow candidates to take units throughout the course of study; allow more unit choice; and include a larger number of types of units. Part of the benefit of using z-scores is that it is able to put a measure on the relative position of how well one candidate does in respect to another taking the same paper. However, this benefit becomes less apparent when the candidates taking any one unit are not the same as those taking another unit.

Both methods suffer from different amounts of over-estimating candidates' marks at the lower end of the mark range and under-estimating candidates' marks at the top end of mark range. The z-score method would not always work in all cases, as it would require a minimum number of candidates entered on a particular unit/component to produce sensible z-scores.

A method to improve on the estimations by effectively applying statistically determined scaling adjustments on the marks to counter the effect of under/over-estimating of marks was suggested. To create these scaling adjustments regression analysis was used. Regression analysis can in its own right estimate marks as it takes into account the correlation between the unit marks. The downside of using this method is that it would require the majority of marks to be available before any estimation of missing marks could take place. Its biggest downfall would most likely be the set-up and processing time required on our exams processing system. Further work using regression analysis to estimate marks is planned.

Overall, it seems both the current method and the proposed z-score method produce similar outcomes for unitised schemes. Most of the new GCSE specifications will be unitised, not linear. Therefore, the benefits of changing the current estimation method do not appear to be that great, and brings into question the amount of effort required to bring in a new method which will make no significant improvement on the current method.

References:

Gray, E. & Shaw, S. (2009). De-mystifying the role of the uniform mark in assessment practice: concepts, confusions and challenges. *Research Matters: A Cambridge Assessment Publication*, **7**, 32–37.

EQUITY ISSUESS

'Happy birthday to you'; but not if it's summertime

Tim Oates Assessment Research & Development, **Dr Elizabeth Sykes** Independent Consultant in Cognitive Assessment, **Dr Joanne Emery, John F. Bell and Dr Carmen Vidal Rodeiro** Research Division

For years, evidence of a birthdate effect has stared out of qualifications data for the United Kingdom; summer-born children appear to be strongly disadvantaged. Whilst those responsible for working on these data have, through mounting concern, periodically tried to bring public attention to this very serious issue, it has been neglected by agencies central to education and training policy. Following a flurry of press interest during 2007 and 2008, it has – justifiably – become a key part of the recommendations which may flow from the Rose Enquiry of the primary curriculum.

Researchers at Cambridge Assessment have had a long interest in the birthdate effect because it is so readily observable in the assessment data that they have worked with (Bell and Daniels, 1990; Massey, Elliott and Ross, 1996; Bell, Massey and Dexter, 1997; Alton and Massey, 1998). More recently, Cambridge Assessment decided to review the issue with the intention to advance the understanding of the extent and causes of the birthdate effect in the English education system (Sykes, Bell and Vidal Rodeiro, 2009). A number of hypotheses have been advanced for its cause – clarity in understanding this fully is a vital part of determining possible remedies. Although the review focuses on understanding the birthdate effect in England, it uses international comparisons as one means of throwing light on key factors.

This article outlines the findings of the review. There is robust evidence from around the world that, on average, the youngest children in their year group at school perform at a lower level than their older classmates (the 'birthdate effect'). This is a general effect found across large groups of pupils. In the UK, where the school year starts on September 1st, the disadvantage is greatest for children born during the summer months (June, July, August). Individual summer-born pupils may be progressing well, but the strength of the effect for the group as a whole is an issue of very significant concern. Since the effect of being the youngest in the year group holds in other countries where the school year begins at other times in the calendar year, medical/seasonality hypotheses regarding pre-natal