CAMBRIDGE ASSESSMENT

Examining the impact of moving to on-screen marking on the stability of centres' results

Tom Benton

Cambridge Assessment Research Report
11th March 2015

University of Cambridge
Local Examinations Syndicate

Author contact details:

Tom Benton
ARD Research Division
Cambridge Assessment
1 Regent Street
Cambridge
CB2 1GG
United Kingdom

benton.t@cambridgeassessment.org.uk

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

How to cite this publication:

Benton, T. (2015). *Examining the impact of moving to on-screen marking on the stability of centres' results*. Cambridge Assessment Research Report. Cambridge, UK: Cambridge Assessment.

Table of Contents

Introduction

In recent years, increasing numbers of examination components have been moved from being marked on paper to be being marked on screen. Within Cambridge Assessment, on-screen marking is done using the Scoris® marking system in partnership with RM plc.

In theory, the move to on-screen marking should provide a number of advantages. A review of some of these potential advantages was provided in Tisi, Whitehouse, Maughan and Burdett (2013, Section 4). One possible advantage, not listed by Tisi et al., is that on-screen marking tends to go hand in hand with a random allocation of the scripts supplied by any centre across different markers. In theory, if some markers are more severe than others, ensuring that no individual centre has all their scripts marked by a particularly severe or lenient marker might reduce instability in centres' results.

Whilst the ideal would be that the effects of unreliable marking are dealt with at root (by ensuring markers themselves more reliable) addressing fluctuations in centres' results due to marking is a worthwhile aim in its own right. The fact that this is a concern amongst schools was highlighted by the Headmasters' and Headmistresses' Conference (HMC) in September 2012 (HMC, 2012) which produced a report criticising the examinations industry for the fact that, amongst other things, there were unexplained fluctuations in the results achieved by particular schools. With this in mind, it is of substantive interest to understand the extent to which marking within Scoris® helps to reduce volatility in centres' results.

This report examines Cambridge Assessment's data across a large number of components to ascertain whether any improvement in year-on-year centre results is associated with moving units to on-screen marking.

Methodology

Within any assessment component[1], the stability in centres' results is quantified by the correlation between the mean raw scores awarded to candidates within each centre in one year and the mean raw scores awarded to candidates in the same centres the following year. A high correlation would indicate that centres that perform well in one year will continue to perform well in the next year. In contrast, a low correlation would imply that the relative performance of centres fluctuates dramatically between years. Correlations in average marks (as opposed to the percentage of candidates achieving particular grades or above across years) are used because they provide a consistent metric across all qualifications regardless of the grading scale that is used. Furthermore, candidates in different centres may be differently distributed across grades so that, whereas in one centre stability in the percentage of candidates achieving C or above is an important metric, in another centre, such as a high-performing selective school, it may be less relevant. Basing our measure of stability upon raw marks allows relevant information to be generated for all assessment components across all centres.

The above measure of stability was calculated for each assessment component of interest. In common with the approach of Benton (2013), the calculation of year-on-year stability in correlations was based upon all centres with at least 20 candidates in each year of interest. Only components with at least 50 such centres were included in the analysis. This means that our analysis is restricted to assessments with relatively large entries. However, concerns over stability in centres' results are most prominent in subjects with large entries and so the restriction to such assessments is of little concern.

We examined the stability in centres' results for every GCSE, A level, AS level, O level[2] and IGCSE component taken in particular annual sessions[3] between 2007 and 2013. For each

[1] Such as a GCSE or GCE unit.

[2] O levels are used within some countries outside of the UK.

component we have also determined the year in which each component was moved to on-screen marking (if at all). Examining these results over time should give us an indication of whether the stability of centres' results has increased or decreased with the move to on-screen marking. More specifically, year-on-year correlations are broken down into four possible groups where:

1. The component is marked on paper in both Year 1 and Year 2
2. The component is marked on paper in Year 1 but on screen in Year 2
3. The component is marked on screen in both Year 1 and Year 2
4. The component is marked on screen in Year 1 but moves back to paper-based marking in Year 2

Note that, of the above scenarios, the fourth is very rare and not considered within this report. Of particular interest, is the change in centre stability that occurs alongside a change in the mode of marking. This is examined by comparing stability between successive pairs of years. Specifically, for each component, we compare:

- Stability between 2007 and 2008 with stability between 2009 and 2010
- Stability between 2008 and 2009 with stability between 2010 and 2011
- Stability between 2009 and 2010 with stability between 2011 and 2012
- Stability between 2010 and 2011 with stability between 2012 and 2013

That is, up to four comparisons of stability for each component.

We can then compare changes in stability where the marking mode has changed to changes in stability where the marking mode has remained constant. Of particular interest are cases where marking is completed on paper in each of the first pair of years but is done on screen for each of the second pair of years. For example, paper-based marking in each of 2007 and 2008 but on-screen marking in each of 2009 and 2010.

Results

The changes in centre-level correlations for each pair of successive sets of two years are shown for each assessment component in Figure 1. As can be seen, components with larger entries (in terms of numbers of common centres[4]) display less change in year-on-year stability than components with smaller entries. The results in Figure 1 are colour coded to distinguish between where:

- Each of the pairs of years is marked entirely on paper (small blue diamonds)
- Each year in the first pair of years is marked on paper but exactly one of the second pair of years is marked on screen (yellow circles)
- Each year in the first pair of years is marked on paper but each year in the second pair of years is marked on screen (large red diamonds)
- Each of the pairs of years is marked entirely on screen (small red triangles).

The (very rare) instances of components moving from on-screen marking back to paper-based marking are not included in Figure 1. Note that, although the same component may occur up to four times in Figure 1 (for the four different pairs of successive years), each component will only be included as having partially or having fully moved to on-screen marking once[5].

The results in Figure 1 clearly show that as components move to on-screen marking the level of stability in year-on-year results tends to increase. In contrast, there is no evidence of a similar

[3] June for GCSE, A level, AS level and IGCSE; November for O level.

[4] That is, the number of centres entering candidates for the component in two successive years.

[5] That is, there is at most one red diamond and one yellow circle for any given component.

improvement where the marking mode has remained stable; either on-screen or paper-based. A summary of the results in Figure 1 is given in Table 1. This table shows that, out of the 35 units where we can track changes in centre-level stability from marking taking place entirely on paper to entirely on screen, almost three-quarters (25 components) display an increase in stability. In contrast, across more than 500 instances where the marking mode is unchanged, only around half show an increase in stability. Furthermore, even when the move to on-screen marking is only partially complete, the majority of instances (27 out of 39) show an increase in centre-level stability.

Figure 1: Changes in year on year centre correlations

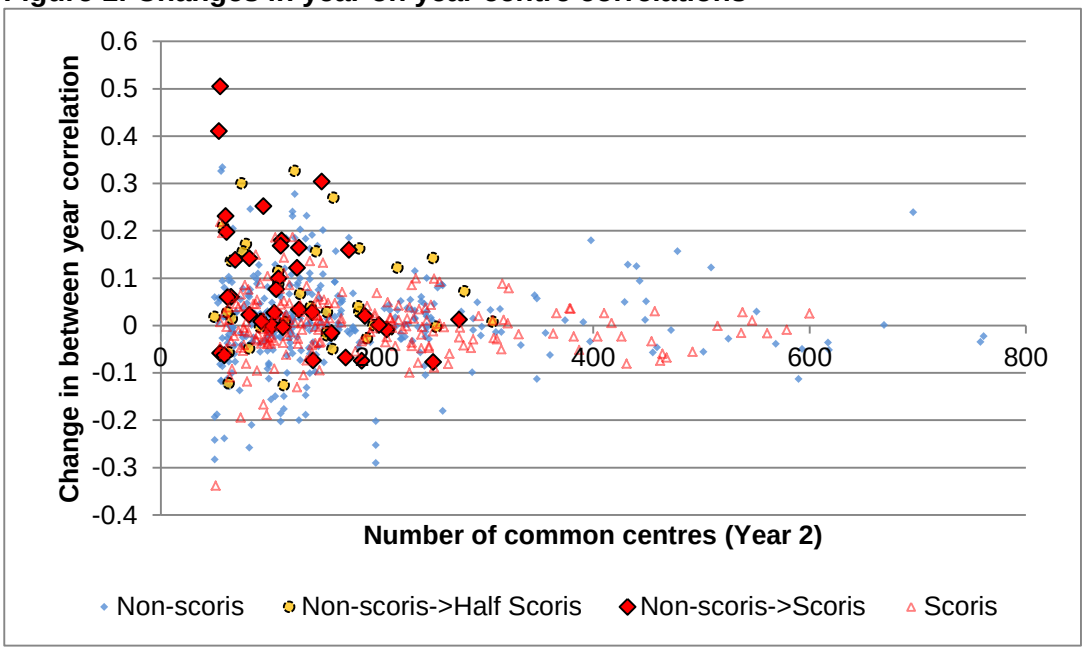


Table 1: Summary of changes in year on year centre correlations

	Change in marking mode			
	Non-scoris->Scoris	**Non-scoris**	**Scoris**	**Non-scoris->Half Scoris**
Average year-on-year correlation for first pair of years (Before)	0.690	0.719	0.763	0.675
Average year-on-year correlation for second pair of years (After)	0.775	0.726	0.763	0.733
Average change in year-on-year centre-level correlations	0.085	0.008	0.000	0.058
Number of instances	35	344	208	39
Number showing increase	25	179	102	27
%age showing increase	71.4%	52.0%	49.0%	69.2%

Figure 1 and Table 1 include results from all components where a sufficient number of large, common centres were available. However, further consideration suggests that the analysis may be irrelevant for a number of these. Specifically, the move to on-screen marking is either irrelevant or unlikely to improve centre-level stability for:

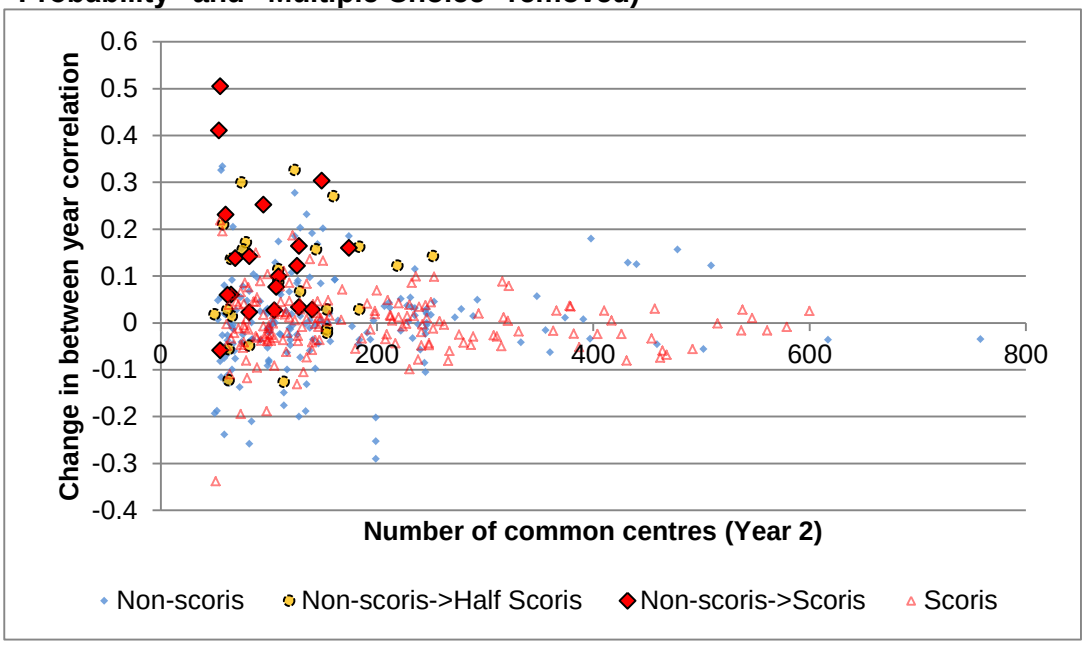
- Components comprising of coursework or other forms of assessments unlikely to be marked on screen at present. This also includes practical assessments in Science.
- Components where marking is either fully objective (such as for multiple choice assessments) or very nearly objective (such as for Mathematics assessments). In these

cases there is little or no possibility of the leniency or severity of individual markers having any effect on centres' results.

With the above thoughts in mind, any components with names containing any of the words "Coursework", "Practical", "Portfolio", "Mathematics", "Mechanics", "Statistics", "Probability" or "Multiple Choice" were removed from the analysis.

The results, once these components are removed, are shown in Figure 2 and summarised in Table 2. As can be seen, the picture is now even more positive than before with 18 out of 19 components that have fully moved to on-screen marking displayed an increase in the stability of centres' results with an average increase in correlation of almost 0.15. Specifically, the average year-on-year correlation in centres' scores increased from 0.60 to 0.74. Assuming no change in the demand of assessments, for a typical component, this would relate to a reduction in the average change in mean scores between years from 3.6 marks to 2.9 marks[6]. Only around half of instances where the marking mode is unchanged show an increase in stability. The differences in results between components moving to on-screen marking and those with a constant marking method are highly unlikely to have occurred by chance alone and strongly suggest that the move to on-screen marking is associated with an increase in the stability of centres' results.

Figure 2: Changes in year-on-year centre correlations (components with the words "Coursework", "Practical", "Portfolio", "Mathematics", "Mechanics", "Statistics", "Probability" and "Multiple Choice" removed)



[6] Calculated as follows:
For the 19 components considered, across all years, the median standard deviation in average scores between centres is 5 marks. This means that, for a typical component, the variance in average scores is 25 marks squared. The correlation between two years is equivalent to the proportion of variance in scores explained by the underlying expected average score within a centre. Thus the residual variance within years is equal to 25*(1-correlation), the variance of the difference between two years is twice this, and, using the properties of the half-normal distribution, the expected size of the change between years is the square root of this times 2/π. Thus:

If a component is marked on paper in both years the expected change is $\sqrt{2 * 25 * (1 - 0.60) * \frac{2}{\pi}} \approx 3.6$.

If a component is marked on screen in both years then the expected change is $\sqrt{2 * 25 * (1 - 0.74) * \frac{2}{\pi}} \approx 2.9$.

Table 2: Summary of changes in year on year centre correlations (coursework, practicals, portfolios, maths and multiple choice assessments removed)

	Change in marking mode			
	Non-scoris->Scoris	**Non-scoris**	**Scoris**	**Non-scoris->Half Scoris**
Average year-on-year correlation for first pair of years (Before)	0.596	0.683	0.760	0.607
Average year-on-year correlation for second pair of years (After)	0.742	0.687	0.760	0.692
Average change in year-on-year centre-level correlations	0.146	0.004	0.000	0.085
Number of instances	19	188	174	26
Number showing increase	18	92	84	20
%age showing increase	94.7%	48.9%	48.3%	76.9%

Discussion

This analysis has revealed that moving components to on-screen marking in Scoris® appears to have a positive influence on the stability of centres' results. Further exploration of individual components revealed no obvious other explanations for this improvement such as changes in the structure or number of marks available on components. As is clear from the analyses, the stability of component results across centres can fluctuate somewhat between years regardless of the form of marking. This means that the increase in stability may not become immediately evident for every individual assessment component that moves to on-screen marking. However, the analysis clearly shows that, on average, a move from paper-based to on-screen marking will tend to increase the stability in centres' results.

It is also worth noting that even after moving to on-screen marking in Scoris®, the level of stability in results need not necessarily be exceptionally high. Previous analysis (Benton, 2013) has shown that, even for assessments with high marking reliability, the achievement of individual centres can fluctuate somewhat over time. Nonetheless, it is clear that a move to on-screen marking and a concomitant random allocation of the scripts from each centre across markers can help to improve the stability of centres' results.

References

Benton, T. (2013). *Formalising and evaluating the benchmark centres methodology for setting GCSE standards*. Cambridge Assessment Research Report. Cambridge, UK: Cambridge Assessment.

Benton, T. (2015). *Examining the impact of moving to on-screen marking on concurrent validity*. Cambridge Assessment Research Report. Cambridge, UK: Cambridge Assessment.

HMC (2012) England's 'examinations industry': deterioration and decay. Downloaded from http://www.hmc.org.uk/publications/.

Tisi, J., Whitehouse, G., Maughan, S., and Burdett, N. (2013). *A review of the literature on marking reliability research*. Coventry: Ofqual.