CAMBRIDGE ASSESSMENT

# *Examining the impact of moving to on-screen marking on concurrent validity*

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/

Cambridge Assessment is the brand name of the University of Cambridge Local Examinations Syndicate, a department of the University of Cambridge. Cambridge Assessment is a not-for-profit organisation.

**How to cite this publication:**

Benton, T. (2015). *Examining the impact of moving to on-screen marking on concurrent validity*. 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). Some of these proposed benefits including continuous marker monitoring, and systems to eliminate errors from incorrect transcription or addition are included within the Scoris® system. Tisi et al. also state that on-screen marking can be beneficial as it allows greater use of item-level rather than whole-script marking.

Ideally, in order to evaluate the move to on-screen marking we would estimate the marking reliability of all components under both modes of marking. However, although it is easy to mass produce marking reliability statistics for components that are marked on-screen (see Dhawan & Bramley, 2012, for examples of making use of data from seed scripts) it is not possible to produce marking reliability statistics for paper-based marking on the same scale. Indeed the recent literature review by Tisi et al. provided very few examples of comparisons of marking reliability between paper-based and on-screen marking (with just some small scale examples provided by Fowles 2008, Johnson et al., 2009, and Johnson et al., 2012).

However, although we are unable to *directly* measure marking reliability before and after components have moved to on-screen marking, we may be able to indirectly infer something about reliability in other ways. Specifically, it is reasonable to assume that if the marking reliability of a component decreases, then its validity must also decrease. Indeed it has often been stated that “reliability is a necessary condition for validity” (Kane, 1992). Thus, we can use information from one of the easily calculable empirical estimators of validity over time to provide evidence about whether reliability has increased or decreased subsequent to a move to on-screen marking.

For the purposes of this report we will focus upon concurrent validity. This report considers concurrent validity in the form of the correlation between scores awarded for a given component and achievement in other examinations taken at the same time. This is preferred to using predictive validity (that is, the ability of scores on one component to predict future achievement) because a greater amount of data is available.

It is worth noting that changes in concurrent validity may be caused by factors other than marking reliability. For example, if the content of an examination became more closely aligned with the measure of concurrent attainment then we may also see increases. Again, it is plausible that if the centres entering candidates for an assessment have a uniform quality of teaching across different subjects/topics we may see greater levels of concurrent validity than if the level of teaching quality is uneven. Thus, for any individual assessment, a change in concurrent validity is not necessarily an indication of a change in marking reliability.

To address the above issue we ensure that analysis considers not just a single assessment but examines changes in concurrent validity across many different assessments. The assumption in such analysis is that, although it is possible that alternative explanations (other than marking reliability) may account for changes in concurrent validity for particular individual assessments, it is unlikely that, on average, such alternative explanations would have more effect on those assessments being moved to on-screen marking than on assessments where the marking mode is unchanged. For example, we are assuming that a move to on-screen marking does not imply that the content of an exam will change to better reflect concurrent validity measures. Again, we are assuming that the move to on-screen marking will have no particular impact on the uniformity of the quality of teaching at participating centres. These would appear to be reasonable assumptions as the move to on-screen marking is not intended to have any impact upon the content or delivery of qualifications, but only upon the way in which they are marked. Provided this assumption holds then, an increase in the *average* level of concurrent validity of qualifications moving to on-screen marking, above and beyond average changes in concurrent validity for those with a constant marking mode, should be indicative of changes in marking reliability.

## Analysis of changes in concurrent validity over time

We examined the concurrent validity of every GCSE, A level and AS level component[1] taken in the June examination sessions between 2007 and 2013. For each 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 marking reliability has increased or decreased with the move to on-screen marking. As well as examining changes in concurrent validity for components that move to on-screen marking, our analysis also estimates changes in concurrent validity for components where the mode of marking remains unchanged; either through marking being paper-based *or* on-screen for two consecutive years. This allows us to identify the extent to which concurrent validity can fluctuate in the absence of any change to the mode of marking.

### *GCSE components*

For each OCR GCSE component we calculated the correlation between the raw score achieved on the component itself and the average GCSE grade achieved in all other GCSEs[2] not including the GCSE towards which the component was contributing. For example, this would include examining the correlation between raw score on the Higher Tier English paper A680 and the average grade candidates achieved on all their other GCSEs excluding English and English Language. Analysis was restricted to components taken in the June session of each year from 2007 to 2013.

In total 807 GCSE components[3] were examined. Of these, 98 components had moved from paper-based to on-screen marking between 2007 and 2013. Of these, well over half (63) displayed an increase in their correlation with concurrent GCSE attainment.

The majority of the 98 changes to on-screen marking happened in 2008 (63), with fewer switches occurring in 2009 (16), 2013 (10), 2010 (6) and 2012 (3). Further details on changes in 2008, 2009 and 2013 are displayed in Figures 1 to 3 against the number of matched candidates for each component (that is, the number with available data on concurrent attainment). As stated earlier, the counterfactual is important and so both components that have changed the mode of marking (red squares) and components where the mode of marking has remained constant (blue diamonds) are included in these charts.

As can be seen from Figure 1, units that moved from paper-based to on-screen marking between June 2007 and June 2008 displayed a greater improvement in concurrent validity than units where the marking mode remained constant[4]. More than three-quarters (48 out of 63) of units increased their concurrent validity with the move to on-screen marking whereas only just over half (109 out of 194) of units where the marking mode was constant increased their concurrent validity over the same period. Thus, Figure 1 appears to indicate that moving to on-screen marking can positively increase the reliability of marking.

[1] Provided at least 500 candidates were available with matching data to the concurrent attainment measure.

[2] Data on achievement in *other* GCSEs was obtained from the National Pupil Database. The National Pupil Database, 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.

[3] Note that these components include coursework and practical units where on-screen marking would not be considered. These are useful as they provide a benchmark against which to compare changes in concurrent validity over time.

[4] Further analysis examined whether components that were consistently marked on-screen displayed greater changes in concurrent validity over time than those consistently marked on paper. No differences were identified.

Figure 1: Changes in concurrent validity of GCSE components between 2007 and 2008

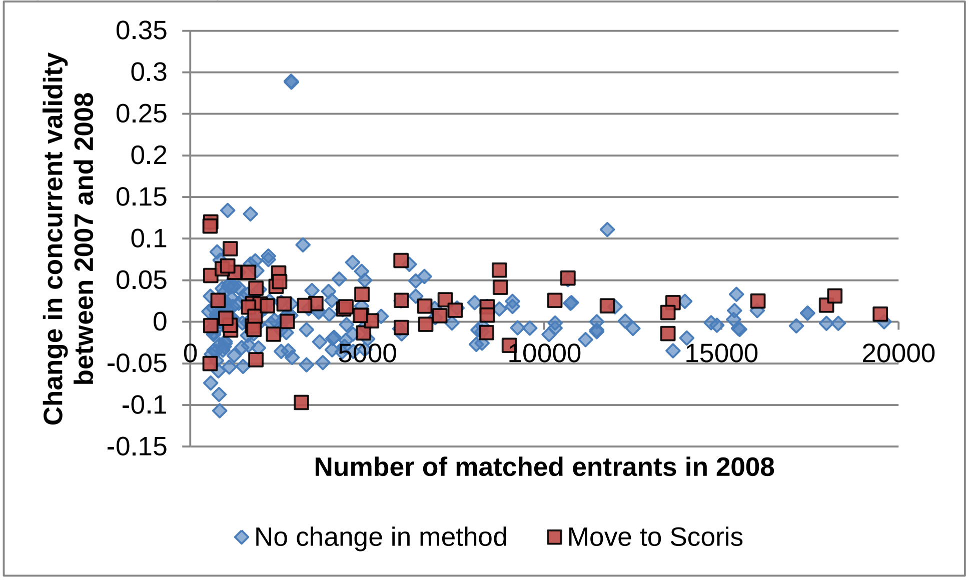


Figure 2: Changes in concurrent validity of GCSE components between 2008 and 2009

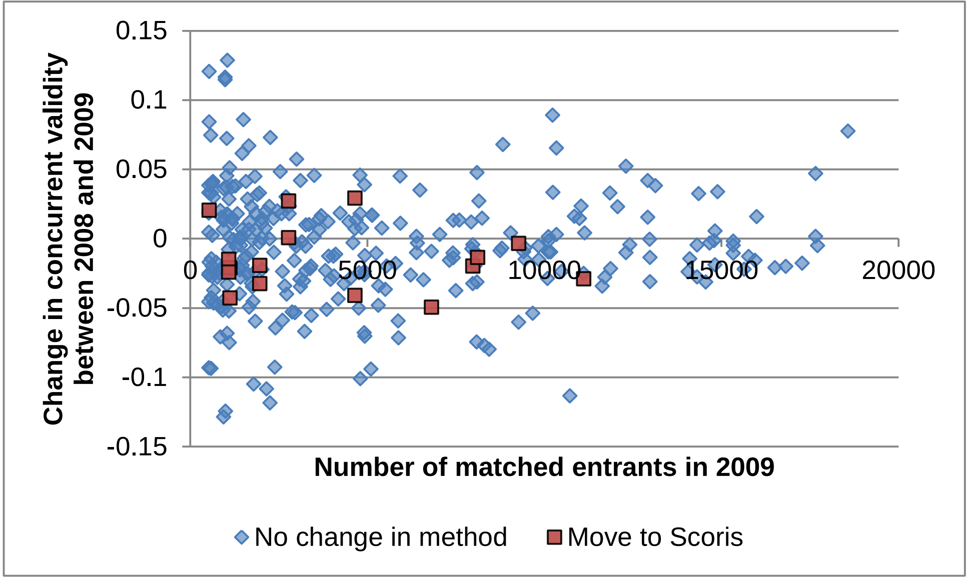
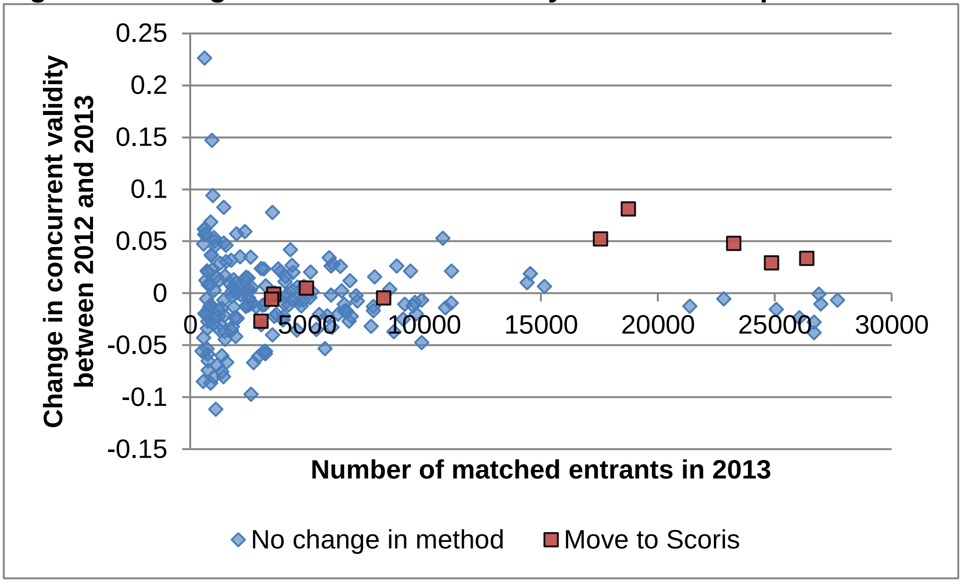


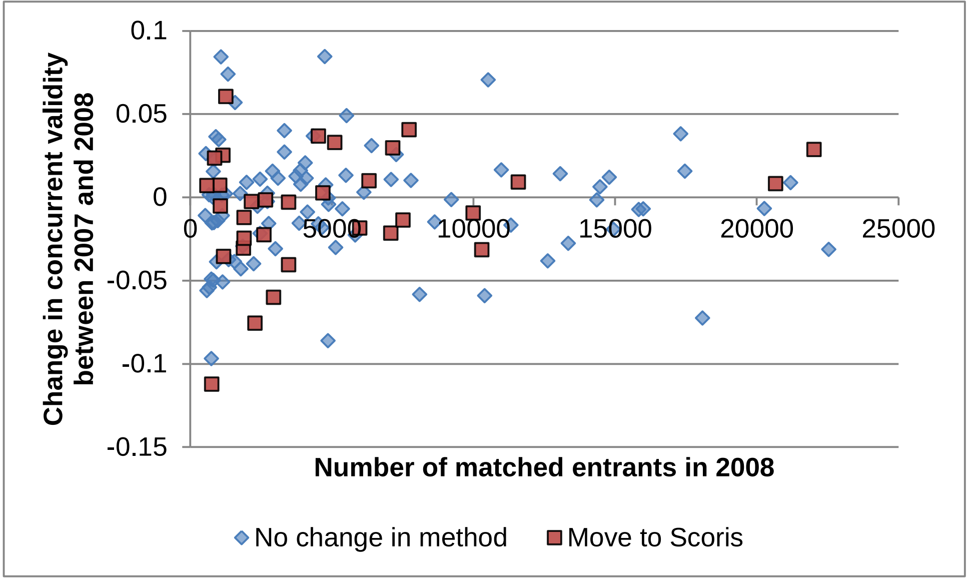
Figure 3: Changes in concurrent validity of GCSE components between 2012 and 2013



Unfortunately, the positive picture presented in Figure 1 is not supported by Figures 2 and 3. Only a minority of those components that moved to on-screen marking in 2009 displayed a simultaneous increase in their concurrent validity. Of those components moving to on-screen marking in 2013, just over half (6 out of 10) displayed a simultaneous increase in concurrent validity. However, further exploration of the units displaying increasing concurrent validity showed that, in part, this was due to increasing heterogeneity in the ability levels of candidates taking these components[5] (evidenced by increased standard deviation of their mean GCSE grade in other subjects). In other cases, the number of marks available on the same units increased[6] providing an alternative explanation for why the concurrent validity may have increased. Having said this, given the general level of variability in the concurrent validity statistics across years (as evidenced by the components that had not moved to on-screen marking) there is no evidence in either Figure 2 or Figure 3 of the move to on-screen marking having had a detrimental impact on reliability.

Given the inconsistency of results, the apparently positive findings in Figure 1 require further scrutiny. Further analysis examined the concurrent validity of each component in Figure 1 from unitised GCSEs in terms of the correlation between the raw score achieved on that component and the total Uniform Mark Scale (UMS) score achieved on the remainder of the units contributing to the same GCSE. The change in concurrent validity (now estimated in this new way) for each component between 2007 and 2008 was estimated both for components that switched to on-screen marking and those that did not. The results of this analysis are shown in Figure 4. As can be seen, the apparently positive results from Figure 1 are not repeated when concurrent validity is calculated in this alternative way, even though many of the exact same GCSE components are being analysed in each case.

**Figure 4: Changes in concurrent validity of GCSE components between 2007 and 2008 (defined via correlation with UMS on remainder of qualification)**



One hypothesis for the differences between the results in Figure 1 and those in Figure 4 is that the nature of the kinds of GCSEs pupils were taking changed between 2007 and 2008 to more closely match the content of components that were moved to on-screen marking. Specifically, we know that many of the components that were moved to on-screen marking in between 2007 and 2008 were those consisting mainly of short answer questions with easily defined correct and incorrect answers. Alongside this we know that the popularity of different GCSEs changed between 2007 and 2008, with particularly large increases in the proportion of pupils taking Separate Sciences rather than combined Science (see Gill, 2013). Thus, the collection of GCSE

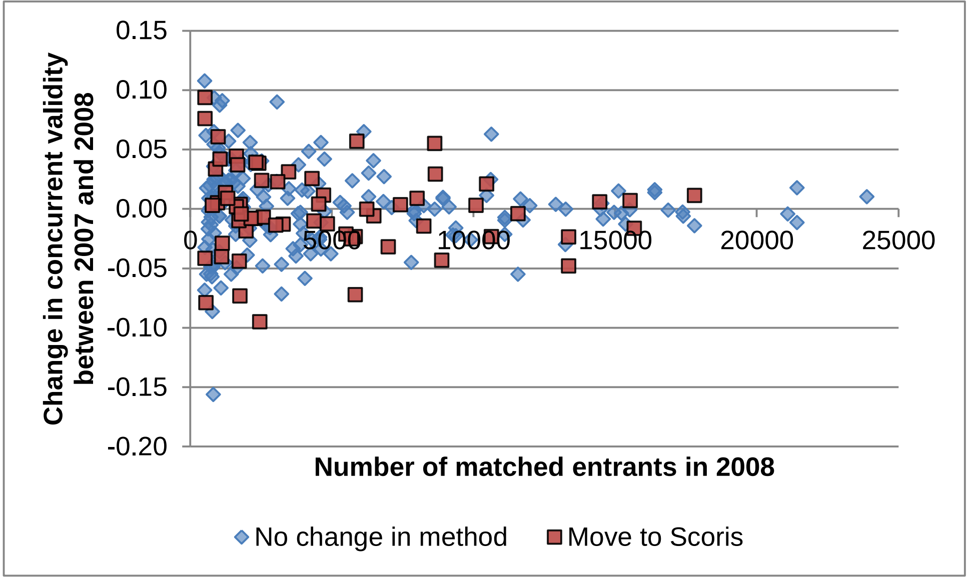
[5] We have verified that this was not the cause of the increase in concurrent validity for components moving to on-screen marking in June 2008.

[6] For example, from 40 to 49 marks on GCSE component A662 or from 50 to 53 marks on A952.

results comprising a pupil's mean GCSE grade would contain a greater proportion of results from Science exams in 2008 than in 2007. This might conceivably imply that the mean GCSE measure was slightly more related to pupils' abilities with regard to short-answer questions in 2008 than in 2007. This might in turn lead to an increase in the correlation with the kinds of components that were switched to on-screen marking.

To examine the above hypothesis further for each component analysed in Figure 1, concurrent validity was estimated in a third way – the correlation between the raw score achieved on each component and the average grade achieved in English and Maths. Using this definition of concurrent validity means that the content of the external measure of ability is fixed between years and should avoid the issue detailed above. Changes in concurrent validity defined in this way between 2007 and 2008 are examined in Figure 5. Analysis using concurrent validity defined in this was shows no major gain between 2007 and 2008 for components that switched to on-screen marking. This implies that our hypothesis above, rather than an increase in the reliability of marking, is likely to explain the results in Figure 1. Nonetheless, in all these analyses, there is no evidence that the switch to on-screen marking has led to a decrease in marking reliability.

**Figure 5: Changes in concurrent validity of GCSE components between 2007 and 2008 (defined via correlation with mean grade on English and Maths GCSEs)**



### *A and AS level components*

For each OCR A level component we calculated the correlation between the raw score achieved on that component and the total UMS score achieved on the remainder of the units contributing to the same unitised A level qualification. For the sake of simplicity, candidates were only included in analysis if they entered the given component and were certificated for the relevant A level in the same session. Only components taken in the June session of each year were included in the analysis.

In total 422 A level components were examined. Of these, 85 components had moved from paper-based to on-screen marking between 2007 and 2013. Of these, just under half (34) displayed an increase in concurrent validity.

The largest number of switches to on-screen marking happened in 2008 (22), although almost as many switches occurred in 2012 (20) and 2013 (16). Further details on changes in 2008, 2012 and 2013 are given in Figures 6 to 8.

**Figure 6: Changes in concurrent validity of A level components between 2007 and 2008**

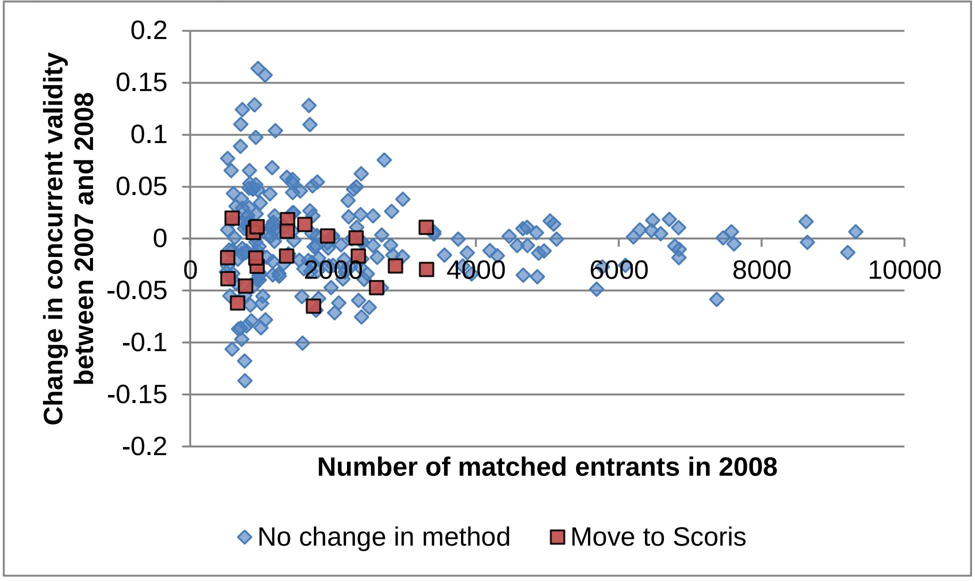


**Figure 7: Changes in concurrent validity of A level components between 2011 and 2012**

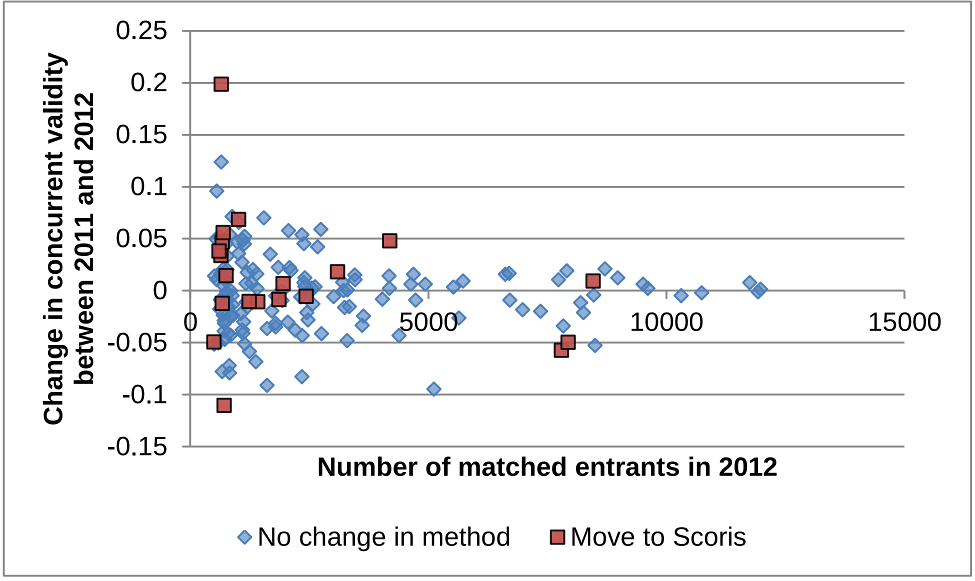
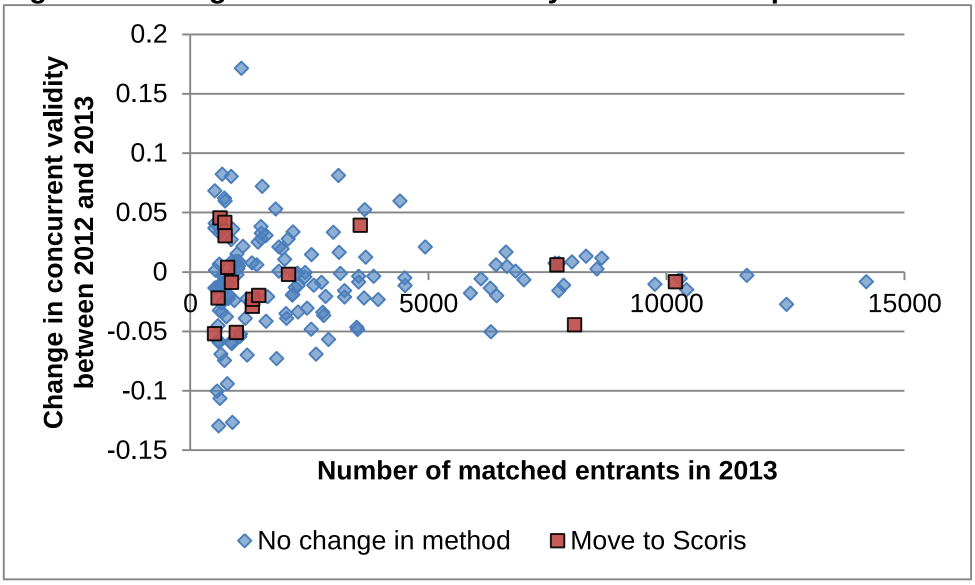


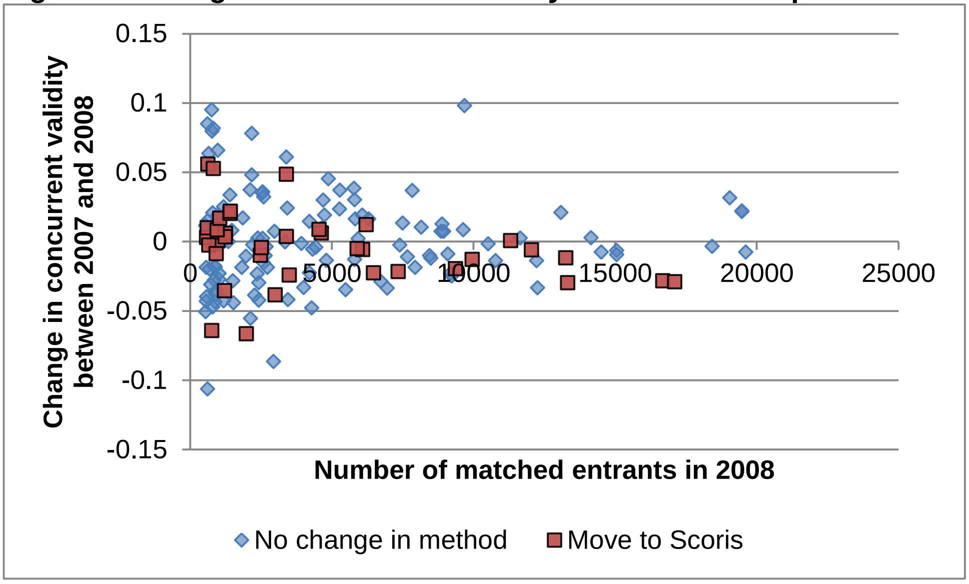
**Figure 8: Changes in concurrent validity of A level components between 2012 and 2013**



Within each of Figures 6, 7 and 8 it can be seen that changes to concurrent validity for units moving to on-screen marking are generally within the same range as changes for units where the marking mode has remained constant. This indicates that, in general, moving units to on-screen marking does not have any impact, either positive or negative, upon the reliability of marking.

Similar analysis to that above was conducted looking at AS levels. In total 80 AS units were examined that moved to on-screen marking between 2007 and 2013. Almost half of these switches (38) occurred in 2008. Figure 9 shows further details of changes in concurrent validity in 2008 for units with different sizes of matched entries. As can be seen no units display a change in concurrent validity that is strongly different to the pattern of changes for units where the marking mode has not changed. This analysis again indicates that, in general, moving paper-based marking to on-screen marking is not associated with a decrease in marking reliability.

**Figure 9: Changes in concurrent validity of AS level components between 2007 and 2008**



## Summary of main findings

This analysis has revealed no evidence of moving components to marking on-screen having any effect on the reliability of marking; either positive or negative. Specifically, it appears that units are at least as likely to display an increase in concurrent validity after moving to marking in Scoris$^{(R)}$ as they are to show a decrease.

As mentioned in the introduction, it is possible that changes in concurrent validity do not necessarily indicate changes in marking reliability. By extension, it might be argued that the lack of change in concurrent validity (on average) seen for the majority of assessments may not indicate a lack of change in marking reliability. However, this report has considered changes in concurrent validity across a large number of assessments. It would seem unlikely that alternative explanations could account for a lack of improvement (on average) in concurrent validity for those units moving to on-screen marking whilst having no effect (on average) on those units where the marking mode remains consistent. As mentioned earlier, a move to on-screen marking is not intended to affect either the content or the delivery of qualifications – only the method of marking should change. For this reason it is unreasonable to dismiss the results in this report on the basis that there may be other explanations for the lack of change.

Having said the above, it remains possible that any *small* improvements in marking reliability would not lead to a detectable improvement in the average level of concurrent validity. Thus it remains possible that the move to on-screen marking may have had a small, beneficial impact on reliability for particular assessments but that we have not been able to detect this having an impact on average concurrent validity. This may be particularly likely for subjects where marking reliability was already high within paper-based marking. After all, if marking reliability was already high, then a change to on-screen marking can only ever lead to small improvements. In this context, it may not be a surprise that little improvement in concurrent validity has been detected alongside the move to on-screen marking.

## References

Dhawan, V. and Bramley, T. (2012). *Estimation of Inter-rater Reliability*. Coventry: Ofqual. http://dera.ioe.ac.uk/17682/1/2013-01-17-ca-estimation-of-inter-rater-reliability-report.pdf.

Fowles, D. (2008). *Does marking images of essays on screen retain marker confidence and reliability?* Paper presented at the International Association for Educational Assessment Annual Conference, September 7–12, Cambridge, UK. http://iaea2008.cambridgeassessment.org.uk/ca/digitalAssets/180469_Fowles.pdf.

Gill, T. (2013). *GCSE Uptake and Results, by Gender 2003-2012*. Statistics Report Series No.66. Cambridge: Cambridge Assessment. http://www.cambridgeassessment.org.uk/Images/150217-gcse-uptake-and-results-by-gender-2003-2012-.pdf.

Johnson, M., Nádas, R., and Bell, J.F. (2009). Marking essays on screen: An investigation intothe reliability of marking extended subjective texts*. British Journal of Educational Technology*, *41*(5), 814–826.

Johnson, M., Hopkin, R., Shiell, H. and Bell, J.F. (2012). 'Extended essay marking on screen: is examiner marking accuracy influenced by marking mode?' *Educational Research and Evaluation*, *18*(2), 107–124.

Kane, M. (1992). An argument-based approach to validity. *Psychological Bulletin*, *112(3),* 527-535.

Tisi, J., Whitehouse, G., Maughan, S., and Burdett, N. (2013). *A review of the literature on marking reliability research (Report for Ofqual)*. Slough: NFER. http://www.nfer.ac.uk/publications/MARK01/MARK01.pdf.