Evidence for the reliability of coursework

Tom Benton

Paper presented at the
17th annual AEA Europe conference
Limassol, Cyprus, 3-5 November 2016

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

benton.t@cambridgeassessment.org.uk
www.cambridgeassessment.org.uk

Abstract

The reliability of an assessment is defined as the extent to which candidates' results would remain stable if the entire assessment exercise was repeated. Whilst numerous studies have evaluated the reliability of written examinations, relatively little has been done to quantify the reliability of internal teacher assessment within schools. This is unfortunate since several high-stakes qualifications in England depend upon the reliability of internal assessment, often in the form of teacher-marked coursework (or similar). The use of coursework is justified in terms of ensuring the validity and authenticity of the learning experience. However, quantitative evidence of reliability is often lacking. This paper will attempt to infer something about the overall reliability of coursework by comparing its value to that of written examinations taken at the same time in predicting future examination scores. This study will use several years of assessment data from General Certificate of Secondary Education (GCSE) and General Certificate of Education Advanced level (GCE A level) assessments in History and English Literature in England. The results will show that coursework is often just as predictive as externally marked tests in forecasting future performance. Since reliability is a necessary pre-condition for (predictive) validity, this suggests that it may be more reliable than is often recognised.

Introduction

The reliability of an assessment is defined as the extent to which the result achieved by any candidate would be repeated if the entire assessment exercise was replicated (Brennan, 2001). In particular, for most item-based examinations we may be interested in how much we might expect a candidate's score to change if they were asked to attempt a different (but equally legitimate) set of questions within an examination and their examination paper was marked by a new marker. Estimating the reliability of such tests is a very well established field of research. Many techniques have been developed for this purpose all of which, in essence, work from the strength of the correlations between the scores on different questions within a test to infer something about the likely correlation of overall test scores with a notional parallel form. Coefficients such as Cronbach's alpha, Guttman's Lambda4 or McDonald's Omega (Revelle & Zinbarg, 2009) are relatively familiar in this context. Such coefficients have been extensively applied to estimate the reliability of written examinations in England (see for example, Bramley & Dhawan, 2012).

Although techniques for estimating reliability are well established for item-based examinations, it has proved much more difficult to estimate the reliability of other types of assessments where item-level data is not available. In some instances this may be because the assessment in question does not consist of items. For example, we may be interested in assessing the reliability of coursework scores. Coursework assessments consist of essentially one overarching task given to all pupils with a single overall mark assigned to their efforts by a teacher. As such, there are no items within the assessment so item-based approaches to reliability are utterly ineffective. Similar considerations may also limit our ability to estimate the reliability of Practical Science tasks, certain vocational assessments and artistic performances. This may explain why there are relatively few studies quantifying the reliability of such assessments including coursework (Johnson, 2012). Indeed, most attempts to estimate reliability for such assessments have focussed purely upon the reliability of marking. Other aspects of reliability, such as what difference it would make if the nature of the task assigned to pupils was changed slightly, are left largely unanswered.

Although it may be difficult to precisely quantify the reliability of coursework, we may be able to infer something about this by looking at how well it predicts future achievement. If it were shown that coursework scores were just as useful in predicting future achievements as scores in formal examinations, then it would be reasonable to assume that the former have a similar level of reliability to the latter. This is the approach that is followed within this paper.

Data analysed

The analysis in this paper looks at the predictive power of coursework and examinations taken in England as part of GCSEs (usually taken by pupils aged 16) in predicting achievement in A level assessments two years later (usually taken by pupils aged 18). The main focus is on the potential of each type of assessment to predict achievement in A level examinations. However, for completeness, it was also of interest to explore the potential of each type of assessment for predicting scores in future coursework assignments. All of the data used for analysis comes from assessments taken as part of GCSEs and A levels offered by the awarding organisation OCR.

Analysis focussed on assessments in English Literature and History. These subjects were chosen as, historically, both of these subjects have involved a substantial amount of coursework both at GCSE and at A level. In addition, they were both taken by large numbers of pupils with OCR at A level meaning that a large amount of data was available for analysis.

For History, analysis focussed upon a GCSE specification that was available until June 2010. Analysis makes use of pupils' coursework scores from this specification as well as their scores on a compulsory written examination that formed part of this GCSE. Scores on the coursework and examination elements at GCSE were linked to coursework and examination[1] scores taken as part of an A level two years later. Separate analyses were conducted based on pupils who took their GCSEs in 2008, 2009 and 2010, and then their A levels in 2010, 2011 and 2012 respectively. Both GCSE assessments were scored out of 50 marks. The A level coursework assessment was scored out of 80 marks and the A level examination out of 120 marks.

For English Literature, analysis used data from GCSE assessments taken slightly more recently as part of a specification first available in June 2012. Due to concerns over workload, and the potential for cheating, tasks equivalent to coursework in all GCSEs at this time were actually taken in the form of controlled assessment (Crisp & Green, 2013). This would mean, for example, that students completed their work within a classroom setting rather than at home. GCSE controlled assessment scores (out of 40) and scores on the most frequently taken GCSE examination (also out of 40) were linked to scores on A level controlled assessment and examinations taken two years later. The analysis examined GCSE assessments taken in 2012 and 2013 linked to A level assessments taken in 2014 and 2015 respectively. It should be noted that, in contrast to History, examinations in GCSE English Literature were tiered meaning that students could choose whether to take a harder (Higher tier) paper where the highest grades were available or a less demanding Foundation tier paper. Since the focus here is on students likely to go on to take an A level, analysis was restricted to students who took the Higher tier examination paper.

In order to simplify the language in the following sections, both controlled assessment and coursework will be referred to under the general heading "coursework".

Correlations with future assessments

Initial analysis in each subject is based upon pupils with data from all four of the relevant assessments: GCSE coursework, GCSE written examination, A level coursework and A level examination. The left hand side of Table 1 shows the inter-component correlations[2] between the different History assessments. Of particular interest are the correlations between GCSE and A level (highlighted in grey), although correlations between assessments within GCSEs and A levels are included for completeness.

[1] There are two options within the A level History examination (Medieval and Early Modern 1066-1715 or Modern 1789-1997). However, the grade boundaries for both options were very similar and for this reason scores from either option were treated interchangeably. Data from both options was used in order to ensure sufficient data was available for analysis.

[2] The correlations in Table 1 are all Pearson correlations. Spearman values were also calculated and were all within 0.02 of the values shown.

In all three years analysed, A level coursework scores are more strongly associated with GCSE coursework than with GCSE examinations. Conversely, in two out of three years, A level examinations are more strongly associated with GCSE examinations than with GCSE coursework. At first glance, this might indicate that coursework and examinations measure separate skills with previous examples of either type of assessment being most relevant to future assessments of the same type.

However, there are two features of the data that are at odds with this neat summary. Firstly, it can be seen in every year that GCSE examinations are more strongly associated with A level coursework than with A level examinations. This implies that, if we wanted to retrospectively estimate a pupil's score on their GCSE exam, their coursework performance at A level would be more useful than their examination score. This would immediately suggest that A level coursework must be at least fairly reliable – if it were not, then why would it be more informative about GCSE exam scores than a separate formal examination?

	All matched pupils				All GCSE pupils		Size of truncation	
	A level coursework (max=80)	A level exam (max=120)	GCSE coursework (max=50)	GCSE exam (max=50)	GCSE coursework (max=50)	GCSE exam (max=50)	GCSE coursework (max=50)	GCSE exam (max=50)
Correlations for the 2008-2010 data set								
A level exam	0.54							
GCSE coursework	0.44	0.34						
GCSE exam	0.41	0.37	0.36		0.62			
Descriptives								
N	3571				49892			
Mean	58.55	77.50	44.50	32.95	39.44	28.34		
SD	11.91	17.57	4.47	5.70	8.25	7.72	54%	74%
Correlations for the 2009-2011 data set								
A level exam	0.51							
GCSE coursework	0.46	0.31						
GCSE exam	0.41	0.31	0.36		0.61			
Descriptives								
N	3277				50224			
Mean	60.56	78.88	44.61	35.35	39.84	30.60		
SD	11.49	15.98	4.27	4.78	8.02	7.16	53%	67%
Correlations for the 2010-2012 data set								
A level exam	0.48							
GCSE coursework	0.45	0.27						
GCSE exam	0.42	0.38	0.32		0.59			
Descriptives								
N	3451				52539			
Mean	59.47	75.99	44.74	34.67	40.30	29.78		
SD	10.84	16.73	4.22	5.07	7.71	7.45	55%	68%

Table 1: Inter-component correlations and descriptive statistics for History GCSE and A level

A second reason for caution in interpreting the correlations above is provided by comparing descriptive information for the data set of matched pupils (i.e., those who did all four relevant assessments), to the full data set of GCSE pupils who took both GCSE assessments of interest (but did not necessarily go on to study History at A level). This descriptive information is provided on the right hand side of Table 1. A visual representation of the same information is given in Figure 1 for the data set of candidates who took their GCSEs in 2008. As can be seen, and as would be expected, candidates who went on to study A level had higher scores on average than the population of GCSE pupils as a whole. However, of more interest is the change in the spread of GCSE scores for the two groups. Both coursework and examinations

scores are less spread out amongst those candidates going on to study A level than amongst the population as a whole. However, this effect is particularly marked for GCSE coursework. As shown in the final two columns of Table 1, in each year, the standard deviation of coursework scores for those going on to A level was only just above half the size of the original standard deviation for the full data set. In contrast, the reduction in the standard deviation of the GCSE examination is much less with between two-thirds and three-quarters of the standard deviation retained amongst the matched data set. It is well known that reducing the standard deviation of any variable will tend to reduce its correlation with other quantities. Thus, the fact that GCSE coursework scores are less strongly correlated with A level exam scores than GCSE exam scores are may simply reflect the greater truncation of coursework scores amongst those pupils who went on to study A level. If students with a greater range of coursework scores had decided to continue to A level, it is reasonable to assume that the correlation between GCSE coursework and A level examinations would increase. An attempt to account for this fact will be provided later in this paper.

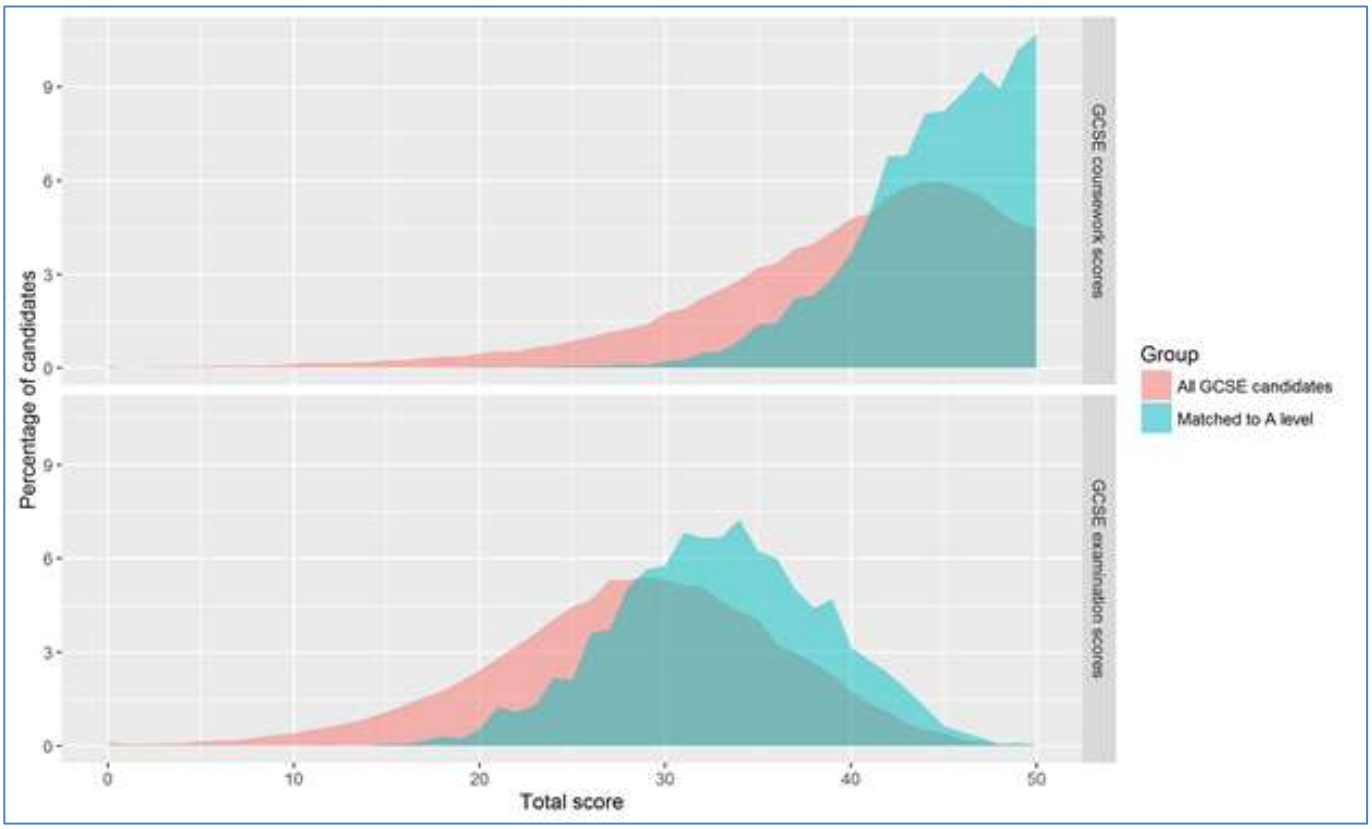

Figure 1: Comparison of score distributions for matched and unmatched pupils for the History 2008-2010 data set

Similar analysis to that in Table 1 is repeated for English Literature in Table 2. Once again, it is clear from the right hand side of Table 2 than GCSE coursework scores in English Literature are truncated to a greater extent amongst those who go on to study A level than GCSE examination scores[3] (also see Figure 2). Again, as before, A level coursework scores are more strongly predicted by GCSE coursework than by GCSE examinations. However, in contrast to History, despite the truncation of the range of GCSE coursework scores, in both years, A level exam scores are more strongly predicted by GCSE coursework than by GCSE examinations. There are only two possible explanations for this finding: that coursework is more relevant to future achievement at A level than GCSE examinations, or that it is measured more reliably. Whichever

[3] In fact the standard deviation of GCSE examination scores is barely lower amongst those going on to study A level than amongst the population as a whole. This may be at least partially explained by the fact that, in contrast to History, English Literature is a tiered examination meaning that the full GCSE population studied as part of Table 2 is already restricted to a higher attaining group of pupils. Note that the heading "All GCSE pupils" refer to all students who took both of the assessments being studied. Specifically this means that the distribution of coursework scores is for those pupils who took the Higher tier examination paper.

of these explanations we prefer (or some combination of the two), it is clear that GCSE coursework scores must provide at least a fairly reliable indicator of pupil ability.

	All matched pupils				All GCSE pupils		Size of truncation	
	A level coursework (max=40)	A level exam (max=60)	GCSE coursework (max=40)	GCSE exam (max=40)	GCSE coursework (max=40)	GCSE exam (max=40)	GCSE coursework (max=40)	GCSE exam (max=40)
Correlations[4] for the 2012-2014 data set								
A level exam	0.59							
GCSE coursework	0.53	0.45						
GCSE exam	0.41	0.40	0.40		0.51			
Descriptives								
N	1523				23527			
Mean	33.62	47.55	35.18	28.31	31.76	24.10		
SD	5.86	9.38	3.82	6.25	5.07	6.87	75%	91%
Correlations for the 2013-2015 data set								
A level exam	0.58							
GCSE coursework	0.57	0.50						
GCSE exam	0.43	0.41	0.47		0.52			
Descriptives								
N	1440				23687			
Mean	33.99	47.62	35.19	27.66	31.55	23.08		
SD	5.67	8.92	3.65	6.41	5.12	7.09	71%	90%

Table 2: Inter-component correlations and descriptive statistics for English Literature GCSE and A level

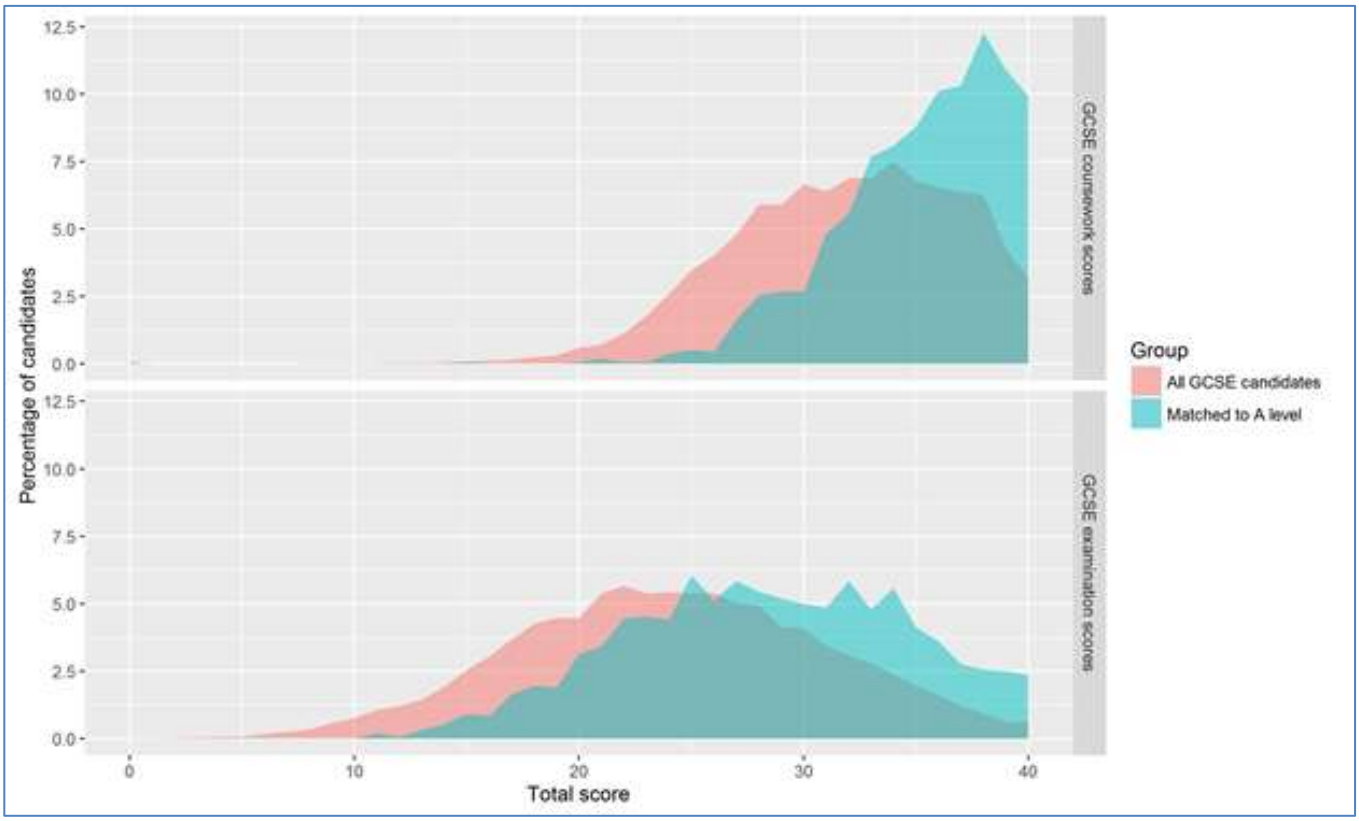

Figure 2: Comparison of score distributions for matched and unmatched pupils for the English Literature 2012-2014 data set

[4] Values in this table are Pearson correlations. Spearman values were also calculated and were very similar with the exception that the correlations between A level coursework and GCSE coursework were roughly 0.05 larger than the values shown in Table 2.

Adjusting for the effect of truncation

As we have mentioned, the correlations in the previous section are affected by the fact that both GCSE scores (but particularly coursework) show less variation amongst those students with matching A level data than amongst the original population of GCSE students. This may lead to a reduction in the apparent predictive validity of coursework relative to examination scores.

There are several existing methods to address the issue of the restricted range of scores available in predictive validity studies (see for example, Wiberg & Sundström, 2009, or Thorndike, 1947). However, many of the existing methods would be inadequate in this case as the restriction of range affects not just one predictor of interest but two and each predictor is affected differently. To address this we use a different methodology based on producing an estimate of what the correlation between GCSE and A level scores would be if all students who had taken the GCSE assessments had gone on to study the A level specification. This method is somewhat speculative in that it relies on extrapolating from the relationship we see for pupils who did study A level to infer something about the relationship for those who did not. Nonetheless, it gives some idea of the possible effect of the restriction of range for coursework scores on the correlations discussed in the previous section.

For the purposes of this section we label the A level score of interest y, and the GCSE coursework and examination scores x_1 and x_2 respectively. Next, we assume a linear regression type relationship between the variables as below.

$$y = \beta_1 x_1 + \beta_2 x_2 + \varepsilon,$$

$$E(\varepsilon) = 0, V(\varepsilon) = \sigma^2$$

We can estimate the parameters of the above linear regression equation (that is, β_1, β_2 and σ^2) using the matched data.

Finally we estimate what the correlation between y and either covariate (before truncation) would be in the full GCSE population using the formula below[5]. To use this formula we substitute in the standard deviations and correlations for the covariates in the data of all GCSE pupils (rather than the matched data only).

$$Cor(y, x_1) = \frac{\beta_1 V(x_1) + \beta_2 Cor(x_2, x_1) SD(x_1) SD(x_2)}{SD(x_1)\sqrt{\beta_1^2 V(x_1) + \beta_2^2 V(x_2) + 2\beta_1\beta_2 Cor(x_2, x_1) SD(x_1) SD(x_2) + V(\varepsilon)}}$$

Table 3 shows the results of this analysis. The final two columns of Table 3 display the estimated correlations between GCSE coursework and examination scores and A level examination scores for the full GCSE populations. The remaining columns show the estimated parameters used in calculations. The standard deviations and correlations between GCSE assessments are estimated based on all pupils who took both and are reproduced from Tables 1 and 2. The regression coefficients are based upon analysis of the matched data set and are shown alongside standard errors[6].

In common with many predictive validity studies we can see that the correlations presented earlier are likely to be an underestimate of the correlation in the population. For example, the

[5] Or a very similar formula if we wish to calculate the correlation between y and x_2.The formula is derived by expanding $Cov(x_1,y)$ and dividing by the standard deviation of x_1 and an expanded version of the standard deviation of y. Note that if $V(x_2)$ is set to zero in this formula then it reduces to an equivalent equation to Thorndike's case 2 formula to adjust for restriction of range (Thorndike, 1947).

[6] Calculated accounting for the clustering of pupils within schools.

relevant correlations in Table 1 (GCSE History) were universally below 0.4, but after correcting for restriction, they are estimated to be above 0.5.

However, of more substantive importance for this paper is the fact that, in all but one of the five cases analysed, scores from coursework appear to be more strongly associated with future achievement than scores from written examinations. Having said this, in two of the instances for History the differences in estimated correlations are very small and are not statistically significant[7]. Nonetheless, the results suggest that we can have at least some confidence in the reliability of this type of assessment.

Subject and A level year	**Results from matched pupils**					**Results from all GCSE pupils**			**Revised correlation of GCSE assessments with A level examinations**	
	β_1	SE	β_2	SE	σ^2	SD(x_1)	SD(x_2)	Cor(x_1,x_2)	Coursework	Written exam
History 2010	0.95	0.11	0.87	0.08	251.24	8.25	7.72	0.62	0.58	0.56
History 2011	0.86	0.09	0.77	0.08	218.67	8.02	7.16	0.61	0.55	0.52
History 2012	0.68	0.10	1.06	0.10	232.79	7.71	7.45	0.59	0.51	0.57
Eng. Lit 2014	0.85	0.06	0.40	0.05	64.67	5.07	6.87	0.51	0.56	0.49
Eng. Lit 2015	0.97	0.07	0.31	0.04	56.68	5.12	7.09	0.52	0.62	0.48

Table 3: Estimated correlation of GCSE coursework or exam scores with A level exam scores after adjusting for truncation

For completeness, Table 4 estimates what the correlation between GCSE assessments and A level coursework would be if the entire GCSE population went on to study A level. As we might expect, the new analysis emphasises the fact that A level coursework scores are more strongly related to coursework scores than examination scores at GCSE. However, more than this, it should be noted that the estimated corrected correlation between GCSE and A level coursework scores are universally higher than any other correlations (corrected or otherwise) calculated as part of this research. It is difficult to see how such high estimates would occur if coursework scores were not at least reasonably reliable.

Subject and A level year	**Results from matched pupils**					**Results from all GCSE pupils**			**Revised correlation of GCSE assessments with A level coursework**	
	β_1	SE	β_2	SE	σ^2	SD(x_1)	SD(x_2)	Cor(x_1,x_2)	Coursework	Written exam
History 2010	0.88	0.07	0.62	0.05	104.08	8.25	7.72	0.62	0.68	0.62
History 2011	0.95	0.07	0.68	0.07	95.27	8.02	7.16	0.61	0.71	0.64
History 2012	0.91	0.07	0.65	0.06	84.02	7.71	7.45	0.59	0.70	0.64
Eng. Lit 2014	0.67	0.04	0.22	0.03	23.04	5.07	6.87	0.51	0.64	0.50
Eng. Lit 2015	0.74	0.05	0.19	0.03	20.52	5.12	7.09	0.52	0.69	0.51

Table 4: Estimated correlation of GCSE coursework or exam scores with A level coursework scores after adjusting for truncation

[7] This can be seen most readily using the fact that, at least within these analyses, the GCSE assessment with the highest estimated correlation is also the one with the highest associated β coefficient. We can see that, for the first two instances in History, the β coefficients differ by less than one standard error so that one coefficient is not significantly higher than the other.

Discussion

This paper has attempted to infer something about the reliability of coursework by looking at the predictive power of scores. The analysis finds that, after adjusting estimates for the restricted range of scores of pupils going to study A level, coursework is often a more powerful predictor of likely success at A level than examination scores. As might be expected, this is particularly true for predicting success in A level coursework but is also sometimes the case for predicting A level examination scores.

There are only two possible explanations for this finding: either in both History and English Literature, coursework scores are often more reliable than those from examinations, or they are more relevant to future achievements. Either way, the results imply that coursework scores must be at least fairly reliable and strengthens the case for coursework as a valid form of assessment.

Of course, demonstrating the reliability of coursework does not in itself prove that it is always the most appropriate form of assessment. For example, demonstrating reliability does not address other concerns over coursework such as the potential for malpractice. Indeed, as an extreme argument (used purely for effect), the high correlation between GCSE and A level coursework scores could be taken to indicate that students who are adept at malpractice, whether this is cheating and/or plagiarising their way to good marks at GCSE, are also effective at employing the same tactics at A level. If this were the case (and I have no evidence to suggest that it is), it could lead to a high correlation between coursework scores without the assessments being in any way valid. Having said this, the fact that coursework scores were often at least as predictive of later examination scores as GCSE exam scores suggests that coursework does indeed reliably measure something of value.

It should be noted that some of the analysis in this report is based upon a somewhat speculative procedure for extrapolating the relationship between GCSE and A level scores from those pupils who have progressed to A level to a wider group of candidates. The rationale for applying this procedure is to address the particularly large reduction in the spread of coursework scores for those pupils going on to study A level. Failing to address this issue is likely to lead to an underestimate of the predictive validity of GCSE coursework scores, meaning that it was important that some adjustment was attempted. Nonetheless, it is true that the precise estimates of corrected correlations provided in Tables 3 and 4 of this paper are dependent on untested assumptions and should be treated with some caution.

Notwithstanding the above criticism, many of the findings in this paper are not dependent upon the procedure used to correct for restriction of range in any case. For example, even before any correction, for both subjects GCSE coursework scores were universally the best predictor of A level coursework scores. Furthermore, for English Literature, A level examination scores were more strongly associated with GCSE coursework than with GCSE exams.

Finally, it should be noted that this research has focussed on two subjects only (History and English Literature). It cannot be assumed that the findings here would generalise to other subjects and further research would be required to verify whether or not this is the case.

References

Bramley, T., & Dhawan, V. (2012) Estimates of reliability of qualifications. In D. Opposs and Q. He (Eds), *Ofqual's Reliability Compendium (pp. 217-320)*. Coventry: Ofqual/12/5117.

Brennan, R. L. (2001). An essay on the history and future of reliability from the perspective of replications. *Journal of Educational Measurement*, *38*, 295–317. http://dx.doi.org/10.1111/j.1745-3984.2001.tb01129.x.

Crisp, V., & Green, S. (2013). Teacher views on the effects of the change from coursework to controlled assessment in GCSEs. *Educational Research and Evaluation*, *19*(8), 680-699. http://dx.doi.org/10.1080/13803611.2013.840244.

Johnson, S. (2012). A focus on teacher assessment reliability in GCSE and GCE. In D. Opposs and Q. He (Eds.), *Ofqual's Reliability Compendium (pp. 365-416)*. Coventry: Ofqual/12/5117.

Revelle, W., & Zinbarg, R. (2009). Coefficients Alpha, Beta, Omega, and the glb: Comments on Sijtsma, *Psychometrika 74* (1), 145–154. http://dx.doi.org/10.1007/s11336-008-9102-z.

Thorndike, R. L. (1947). *Research problems and techniques. (Report No. 3)*. Washington DC: U.S. Government Printing Office.

Wiberg, M. & Sundström, A. (2009). A comparison of two approaches to correction of restriction of range in correlation analysis. *Practical Assessment Research & Evaluation*, *14*(5). Available at http://pareonline.net/getvn.asp?v=14&n=5.