checks (details on the quality assurance procedures can be found in QCA Code of Practice, 2005). In this research we examined the marks without performing these procedures.

References

Bramley, T. (2007). Quantifying marker agreement: terminology, statistics and issues. *Research Matters: A Cambridge Assessment Publication*, **4**, 22–28.

Brooks, V. (2004). Double marking revisited. *British Journal of Educational Studies*, **52**, 29-46.

Cannings R., Hawthorne K., Hood K. & Houston H. (2005). Putting double marking to the test: a framework to assess if it is worth the trouble. *Medical Education*, **39**, 299-308.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, **20**, 7-46.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, **70**, 213- 220.

Cresswell M.J. (1983). *Optimum weighting for double marking procedures*. AEB Research Report, RAC281.

Landis, J.R. & Koch, G.G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, **33**, 159-174.

Lucas, A.M. (1971). Multiple marking of a matriculation Biology essay question. *British Journal of Educational Psychology*, **41**, 78-84.

Massey, A. & Foulkes, J. (1994). Audit of the 1993 KS3 Science national test pilot and the concept of quasi-reconciliation. *Evaluation and Research in Education*, **8**, 119-132.

Newton, P. (1996). The reliability of marking of General Certificate of Secondary Education scripts: Mathematics and English. *British Educational Research Journal*, **22**, 405-420.

Partington, J. (1994). Double marking students' work. *Assessment and Evaluation in Higher Education*, **19**, 57-60.

Pilliner, A.E.G. (1969). Multiple marking: Wiseman or Cox? *British Journal of Educational Psychology*, **39**, 313-315.

QCA (2005). *GCSE, GCE, VCE, GNVQ and AEA Code of Practice 2005/2006*. London: Qualifications and Curriculum Authority.

Uebersax, J. (2003). *Statistical Methods for the Analysis of Agreement*. http://ourworld.compuserve.com/homepages/jsuebersax/agree.htm. Accessed April 2006.

White, R. (2001). Double marking versus monitoring examinations. *Philosophical and Religious Studies*, **1**, 52-60.

Wilmut, J. (1984). *A pilot study of the effects of complete or partial removal of marks and comments from scripts before re-marking them*. AEB Research Report RAC315.

Wood, R. & Quinn, B. (1976). Double impression marking of English language essays and summary questions. *Educational Review*, **28**, 229-246.

Item-level examiner agreement

Nicholas Raikes and Alf Massey Research Division

Abstract

Studies of inter-examiner reliability in GCSE and A-level examinations have been reported in the literature, but typically these focused on paper totals, rather than item marks. See, for example, Newton (1996). Advances in technology, however, mean that increasingly candidates' scripts are being split by item for marking, and the item-level marks are routinely collected. In these circumstances there is increased interest in investigating the extent to which different examiners agree at item level, and the extent to which this varies according to the nature of the item.

Here we report and comment on intraclass correlations between examiners marking sample items taken from GCE A-level and IGCSE examinations in a range of subjects.

The article is based on a paper presented at the 2006 Annual Conference of the British Educational Research Association (Massey and Raikes, 2006).

Introduction

One important contribution to the reliability of examination marks is the extent to which different examiners' marks agree when the examiners mark the same material. Without high levels of inter-examiner agreement, validity is compromised, since the same mark from different examiners cannot be assumed to mean the same thing. Although high reliability is not a sufficient condition for validity, the reliability of a set of marks limits their validity.

Research studies have in the past investigated inter-examiner reliability, but typically these focussed on agreement at the level of the total mark given to scripts. The operational procedures followed by examination Boards for documenting examiner performance also often involve recording details of discrepancies between examiners at the script total level. New technologies are facilitating new ways of working with examination scripts, however. Paper scripts can now be scanned and the images transmitted via a secure Internet link to examiners working on a computer at home. Such innovations are creating an explosion in the amount of item-level marks available for analysis, and this is fostering an interest in the degree of inter-examiner agreement that should be expected at item level. The present article provides data that will help inform discussions of this issue.

The source of our data

The analysis presented in the present article was of data collected during trials of new ways for examiners to record item-level marks. All marking for the trials was done using paper scripts (i.e. no marking was done on screen, the only innovation was in the way the markers recorded their marks). The marks therefore give an indication of the kind of agreement that can be expected between examiners marking whole scripts on paper. The results are indicative only because the study marking was low stakes for the examiners (i.e. no candidate's result depended on the marks and the examiners knew their performance would not be appraised), and also because different methods of recording marks were being trialled, which might have had a small effect on their reliability.

The five components for which data were available are as follows:

- **IGCSE Foreign Language French: Listening**
  Multiple choice (m/c) and short answer textual answers worth 1 or 2 marks
- **IGCSE Development Studies: Alternative to Coursework**
  Short answers worth 1–6 marks
- **A-level Chemistry: Structured Questions**
  m/c and short answers worth 1–5 marks
- **A-level Economics: Data Response and Case Study**
  Short, textual answers worth 1–6 marks; some longer textual answers worth 8–12 marks
- **A-level Sociology: Principles and Methods**
  Candidates chose 2 from 6, 25-mark essay items

Inter-examiner agreement at script-total level

Although item-level data are the main focus of our article we present results for script totals in Table 1. 'ITR per mark' in Table 1 is the Implied Time Restriction per mark, equal to the time allowed for the examination divided by the maximum mark available for the examination. The column labelled 'ICC r_{totals}' gives the intraclass correlation coefficient between the examiners' total marks for the scripts. The intraclass correlation may be interpreted as the proportion of variance in the set of marks that is due to the candidates (i.e. after examiner effects have been controlled for). That is, if there is perfect agreement between the examiners on every script, the intraclass correlation coefficient will be 1; but if there is no agreement and the marks appear random, the coefficient will be 0. Bramley (2007) discusses approaches to quantifying agreement between pairs of examiners in this issue of *Research Matters*, but correlation based measures are useful when considering the relationship between more than two examiners, as is the case here.

Table 1 : Intraclass correlations for script totals

Subject	*Max mark*	*Time (mins)*	*ITR per mark*	$N_{examiners}$	$N_{scripts}$	*ICC* r_{totals}
French	48	45	0.9	4	300	0.995
Dev. Stud.	35	90	2.6	4	265	0.917
Chemistry	60	60	1.0	3*	298	0.992
Economics	50	110	2.1	4	294	0.774
Sociology	50	90	1.8	3*	252	0.863

* One Chemistry and one Sociology examiner dropped out of the trials

Looking first at the Implied Time Restrictions per mark in Table 1, it seems that the question paper designers generally gave candidates about 1 minute per mark for papers consisting of multiple choice and short answer questions, and about 2 minutes per mark for papers involving more extended answers. Development Studies was apparently generous in the amount of time given to candidates, since this question paper only contained short answer questions.

All the ICCs in Table 1 are high, indicating a considerable degree of agreement between the examiners. As might be expected, the agreement was highest for the French and Chemistry papers, consisting of multiple choice and short answer questions, a little lower for Development Studies, containing only short answer questions, and a little lower still for Sociology, consisting solely of 25-mark essays. It is slightly surprising that the Economics examiners showed the lowest levels of agreement, given that the Economics paper contained some short answer questions. However, as discussed below, the ICC for Economics does not appear low when the Implied Time Restriction is taken into account.

There is a striking relationship between the Implied Time Restriction per mark and ICC. If Development Studies with its apparently generous time restriction is excluded, the Pearson correlation between these two quantities is -0.99 – that is, the degree of agreement between examiners at script-total level for these four question papers can be almost perfectly predicted from the Implied Time Restriction per mark.

Inter-examiner agreement at item level

We classified items as 'objective', 'points' or 'levels' according to the kind of marking required as follows:

- **Objective marking** – items that are objectively marked require very brief responses and greatly constrain how candidates must respond. Examples include items requiring candidates to make a selection (e.g. multiple choice items), or to sequence given information, or to match given information according to some given criteria, or to locate or identify a piece of information (e.g. by marking a feature on a given diagram), or to write a single word or give a single numerical answer. The hallmark of objective items is that all credit-worthy responses can be sufficiently pre-determined to form a mark scheme that removes all but the most superficial of judgements from the marker.
- **Points based marking** – these items generally require brief responses ranging in length from a few words to one or two paragraphs, or a diagram or graph, etc. The key feature is that the salient points of all or most credit-worthy responses may be pre-determined to form a largely prescriptive mark scheme, but one that leaves markers to locate the relevant elements and identify all variations that deserve credit. There is generally a one-to-one correspondence between salient points and marks.
- **Levels based marking** – often these items require longer answers, ranging from one or two paragraphs to multi-page essays or other extended responses. The mark scheme describes a number of levels of response, each of which is associated with a band of one or more marks. Examiners apply a principle of best fit when deciding the mark for a response.

Tables 2 to 4 present data about inter-examiner agreement at item level. Looking first at the bottom right hand cell of each table, the overall mean

Table 2 : means and standard deviations of ICCs for OBJECTIVE items

	Mean ICC (Objective items)				
	(N_{items}) *Standard Deviation of the ICCs*				
Max mark	*French*	*Dev. Stud.*	*Chemistry*	*Economics*	*All*
1	**0.975** (21) *0.027*	**0.981** (1)	**0.950** (3) *0.073*	**0.978** (1)	**0.972** (26) *0.033*
2	**-**	**0.978** (1)	**-** -	**-** -	**0.978** (1)
6	**0.986** (1)	**-**	**-**	**-**	**0.986** (1)
All	**0.975** (22) *0.027*	**0.980** (2) *0.002*	**0.950** (3) *0.073*	**0.978** (1)	**0.973** (28) *0.032*

Table 3 : means and standard deviations of ICCs for POINTS items

	Mean ICC (Points items)				
	(N_{items}) *Standard Deviation of the ICCs*				
Max mark	*French*	*Dev. Stud.*	*Chemistry*	*Economics*	*All*
1	**0.877** (15) *0.082*	**0.883** (2) *0.044*	**0.837** (25) *0.090*	**-**	**0.854** (42) *0.086*
2	**0.852** (3) *0.058*	**0.609** (4) *0.156*	**0.885** (12) *0.062*	**0.774** (2) *0.149*	**0.817** (21) *0.138*
3	**-**	**0.719** (4) *0.082*	**0.899** (3) *0.049*	**0.517** (3) *0.034*	**0.712** (10) *0.165*
6	**-**	**0.809** (1)	**-**	**0.548** (1)	**0.679** (2) *0.185*
All	**0.873** (18) *0.078*	**0.717** (11) *0.143*	**0.856** (40) *0.082*	**0.608** (6) *0.147*	**0.820** (75) *0.126*

ICC is, as expected, highest for the objective items (0.973), next highest for the points items (0.820), and lowest for the levels items (0.773).

Table 2 shows the objective items were marked very reliably regardless of the subject or the maximum mark available (though most of the objective items were on the French Listening paper and only two were worth more than one mark).

One-mark points items (top row of Table 3) were marked a little less reliably than one-mark objective items (top row of Table 2), as expected. The right-most column of Table 3 shows that overall, mean ICC for the points items decreased with rising maximum mark. Surprisingly, this trend does not apply within all the subjects. For Chemistry, the only subject with a considerable number of items worth more than one mark, there is a rising trend.

The six 25-mark Sociology essay items (near the bottom right of Table 4) marked using a levels marking scheme were marked very reliably (average ICC = 0.825, with little variation between the items). It is not

Table 4 : means and standard deviations of ICCs for LEVELS items

	Mean ICC (Levels items)				
	(N_{items}) *Standard Deviation of the ICCs*				
Max mark	*Dev. Stud.*	*Chemistry*	*Economics*	*All*
4	**0.890** (1)	**-**	**-**	**0.890** (1)
8	**-** (1)	**0.740**	**-**	**0.740** (1)
10	**-** (1)	**0.567**	**-**	**0.567** (1)
12	**-** (1)	**0.585**	**-**	**0.585** (1)
25	**-**	**-**	**0.825** (6) *0.044*	**0.825** (6) *0.044*
All	**0.890** (1)	**0.631** (3) *0.095*	**0.825** (6) *0.044*	**0.773** (10) *0.115*

obvious why there was less inter-examiner agreement for the Economics levels items, though the Economics examiners also had the lowest overall mean ICC for the points items. The Sociology results show it is possible to have lengthy pieces of extended writing marked reliably.

Conclusion

In this article we have provided some detailed information about inter-examiner agreement levels that were obtained from IGCSE and A-level examiners marking whole scripts on paper in a non-live context from examinations in five subjects.

Intraclass correlation (ICC) coefficients generally indicated a high degree of agreement between examiners at both script total and item level. When items were classified according to their marking schemes as 'objective', 'points' or 'levels', the objective items were on average marked more reliably than the points items, which were on average marked more reliably than the levels items, as expected. On average reliability decreased with rising maximum mark for points items, but surprisingly this trend was reversed for Chemistry. Six 25-mark Sociology essay questions marked against a levels mark scheme were marked very reliably, proving that it is possible to achieve high reliability for essay marking.

We found a very strong relationship between the Implied Time Restriction (ITR) per mark[1] that was imposed on candidates and the intraclass correlation (ICC) obtained for script total marks. A Pearson correlation of -0.99 was found between ITR per mark and ICC when one subject, IGCSE Development Studies, which had an apparently long ITR per mark, was excluded from the calculation. Implied Time Restriction per mark therefore appears to be a useful indicator of the level of inter-examiner agreement that should be expected at total script mark level.

1 The Implied Time Restriction per mark equals the time allowed for an examination divided by the maximum mark available for the examination, i.e. it is the average time available to candidates for earning a mark.

References

Bramley, T. (2007). Quantifying marker agreement; terminology, statistics and issues. *Research Matters: A Cambridge Assessment Publication*, **4**, 22–28.

Massey, A.J. & Raikes, N. (2006). *Item-Level Examiner Agreement*. Paper presented at the 2006 Annual Conference of the British Educational Research Association, 6–9 September 2006, University of Warwick, UK.

Newton, P.E. (1996). The Reliability of Marking of General Certificate of Secondary Education Scripts: Mathematics and English. *British Educational Research Journal*, **22**, 4, 405–420.

Fostering communities of practice in examining

Andrew Watts Cambridge Assessment Network

This is a shortened version of a paper given at the International Association for Educational Assessment Conference in May 2006.

The necessity of communities of practice in a judgemental system

The term 'community of practice', when applied to examining in a traditional system, is usually used to denote the system of induction, cooperative working, supervision and development of examiners that aims to overcome the error to which their judgements are prone. Dylan Wiliam wrote in 1996 that 'maintenance of standards requires that those responsible for setting standards are full participants in a community of practice, and are trusted by the users of assessment results'. His observation does not only apply to assessments of school attainment. Alison Wolf (1995), writing about competence-based assessment, describes how assessors 'operate in terms of an internalised, holistic set of concepts'. With examples from a number of educational and vocational contexts she concludes '… how important and, potentially, how effective assessor networks are. They are, in fact, the key element in ensuring consistency of judgement' (p.77).

Subjectivity and objectivity

It has been common to characterise the judgements made in assessment as 'subjective' in contrast to more automated assessments which are 'objective'. Pierre Bourdieu (1990) however, in his analyses of social practice, calls any division between these two concepts 'artificial' and particularly argues against the privileging of an 'objective' standpoint. Sueellen Shay (2005) applies Bourdieu's analysis to the case of a university Engineering Department's assessment of undergraduates' final year theses, which she describes as 'complex tasks'. She describes such assessments within the logic of social practice and asserts that 'all judgement is both objectively and subjectively constituted'. She writes that this kind of professional judgement requires 'a double reading … an iterative movement'. From an objective perspective, assessors can 'observe, measure and map reality independent of the representations of those who live in it'. Subjectively, on the other hand, assessment is 'an embodiment of the assessor'; it is 'relational', 'situational', 'pragmatic' and 'sensitive to the consequences of [the] assessment'. Such 'double readings' enable the judges to assess a 'socially constituted, practical mastery' (p.675).

Shay's concept of a socially based 'double reading' presents us with a *requirement* for assessment to take place within a community of practice. Thus, assessment is understood within a social theory of learning, such as Wenger's (1998), which recognises the place of components like 'community, identity, meaning and practice' (p.5). This supports the view that a balancing of subjective and objective perspectives should be sought in making assessments, and that the community of practice provides an appropriate context for the assessment of complex tasks.

Reliability and the use of new technologies

Concern for greater reliability has motivated the search for more automated ways of managing and marking examination scripts. Paper scripts can be scanned and the images transmitted via a secure Internet connection to markers working on a computer at home. There is then the potential for all examiners to mark the same training scripts online, and for a Team Leader to call up instantly any script that an examiner wishes to discuss with them. Team Leaders may more closely monitor and support examiners during marking, since all marked scripts, together with the marks and annotations, are instantly available. Standardising scripts, with marks already agreed by senior examiners, can be introduced 'blind' into online marking allocations to check that examiners have not drifted from the common standard, and statistical methods for flagging potential aberrant marking may be employed. All these procedures may improve the reliability of marking, but they might also undermine the argument for maintaining a community of practice amongst all examiners. If the bulk of examiners can be trained and effectively monitored online, do they need to come together at all?