An Application of Judgement Analysis to Examination Marking in Psychology

James Elander and David Hardman
London Guildhall University, London, UK

Correspondence: j.elander@derby.ac.uk

Cite as:
Elander, J. & Hardman, D. (2002). An application of judgement analysis to examination marking in psychology. *British Journal of Psychology*, 93, 303-328. DOI: 10.1348/000712602760146233

Abstract

Statistical combinations of specific measures have been shown to be superior to expert judgement in several fields. In this study judgement analysis was applied to examination marking to investigate factors that influenced marks awarded and contributed to differences between first and second markers. Seven markers in psychology rated 551 examination answers on seven 'aspects' for which specific assessment criteria had been developed to support good practice in assessment. The aspects were addressing the question, covering the area, understanding, evaluation, development of argument, structure and organisation, and clarity. Principal components analysis indicated one major factor and no more than two minor factors underlying the seven aspects. Aspect ratings were used to predict overall marks, using multiple regression regression to 'capture' the marking policies of individual markers. These varied from marker to marker in terms of the numbers of aspect ratings that made independent contributions to the prediction of overall marks and the extent to which aspect ratings explained the variance in overall marks. The number of independently predictive aspect ratings, and the amount of variance in overall marks explained by aspect ratings, were consistently higher for first markers (question setters) than for second markers. Co-markers' overall marks were then used as an external criterion to test the extent to which a simple model consisting of the sum of the aspect ratings improved on overall marks in the prediction of co-markers marks. The model significantly increased the variance in co-markers' marks accounted for, but only for second markers, who had not taught the material and not set the question. Further research is needed to develop the criteria and especially to establish the reliability and validity of specific aspects of assessment. The present results support the view that, for second markers at least, combined measures of specific aspects of examination answers may help to improve the reliability of marking.

Introduction

This paper considers the performance of markers in psychology examinations from the perspective of the psychology of expert judgement. The task of a marker involves making an overall assessment of the quality of an answer, taking into account a number of more specific features or aspects, such as the accuracy and completeness of the material and the level of argument and critical evaluation. This makes the task potentially amenable to the type of analysis that has been applied to a wide range of situations involving expert judgment and decision making.

Einhorn (2000) considered the tasks that must be undertaken by an expert making judgments. One is to *identify* information, or cues, from the multidimensional stimuli they encounter. A second is to *measure* the amount of the cue. A third is to *cluster* the information into smaller numbers of dimensions. When these three tasks have been achieved, an overall evaluation can be made by weighting and combining the cues. It is this integration of information about multiple cues that research has shown human experts have the most difficulty with. "People are good at picking out the right predictor variables and coding them in such a way that they have a conditionally monotone relationship with the criterion. People are bad at integrating information" (Dawes, 1982, p. 395).

That view reflects the results of a very large body of findings where the statistical combination of separate items of information was shown to be superior to a single overall judgement. The judgements that have been examined in this way include clinical assessment (Goldberg, 1968; Goldman et al., 1988; Leli & Filskow, 1981), student selection (deVaul et al., 1957), parole board decisions (Carrol et al., 1988) and the prediction of business failure (Deacon, 1972). One of the most compelling examples of expert judgement being outperformed by statistical methods was where information from three pathologists' examinations of biopsy slides was used to predict survival time for patients with Hodgkin's disease (Einhorn, 1972). The pathologists' overall ratings of disease severity were not related to survival times, but statistical combinations of their ratings of nine histological characteristics of each slide were. The prediction of survival time by the nine components of the statistical model was compared with the components plus the overall ratings of severity. Those results differed from judge to judge, with one judge appearing to benefit from the addition of the overall severity rating to the model but not the other two. Einhorn concluded: "It seems that in certain cases the global judgement does add to the components and should be included in the prediction equation, while in other cases its inclusion only tends to lower the probability. This is obviously an empirical question that can only be answered by doing the research in the particular situation" (Einhorn, 1972, p. 96).

The most commonly used statistical method for combining separate items of information is multiple regression, where predictor variables are weighted in such a way as to maximise the correlation between the subsequent weighted composite and the target variable (Cooksey, 1996, chapter 4). This approach can also be used to 'capture' the judgement policy of an individual expert or group of experts, by identifying the weight that is attached to different items, or cues, in the making of an overall judgement. The approach is often traced back to Hoffman's (1960) derivation of the concept of relative weight as an appropriate way of representing the judgement processes of an individual. This approach does not claim to provide a complete description of the judgement process, but is described as a "paramorphic mathematical representation that "captures" aspects of the judgement process" (Cooksey, 1996, p. 157). Regression methods where weightings are assigned to predictor variables are used in what Dawes (1982) called “proper linear models”. “Improper linear models,” by contrast, involve combinations of predictors that are not weighted, or weighted in a sub-optimal way, and are appropriate in situations where there is no clear criterion for the judgements being made (Dawes, 1982).

In one review of the evidence from studies that compared expert judgement with statistical methods, Dawes observed that “in both the medical and business contexts, exceptions to the general superiority of actuarial judgement are found where clinical judges have access to more information than the statistical formulas used.’ (Dawes, 1994, p. 93). In student assessment, access to and use of information unrelated to the criteria for assessment is exactly what one wishes to

avoid. Dennis et al. (1996), for example, found that much of the variance in supervisors' marks of psychology student projects was attributable to influences specific to the supervisor, possibly reflecting personal knowledge of the student. Blind marking can reduce such biases in examinations, but not necessarily eliminate them; some markers may, for example, recognise the handwriting of individual students.

Judgement analysis has not to our knowledge been applied to examination marking, but could potentially lead to the development of more reliable assessments, as well as providing insights into the ways that markers arrive at overall judgements about students' work and the sources of discrepancies between markers. This would depend, however, on the identification of the relevant 'cues' for the assessment of students' work and their accurate measurement.

Educational research on essay marking in psychology has already gone some way towards specifying candidate 'cues' or aspects of assessment. Norton (1990) conducted detailed interviews with coursework essay markers in psychology about the things they looked for when marking and what they considered important. There was considerable variation in responses, with 18 different aspects nominated, many of which were overlapping in meaning. The nine aspects mentioned by at least half of the tutors were structure, argument, answering the question, wide reading, content, clear expression of ideas, relevant information, understanding, and presentation (Norton, 1990, table 13). The interview data revealed how variable and subjective the marking process can be, and Norton concluded: "On the one hand there was a remarkable consistency about the central importance of argument, structure and relevance. On the other hand there were quite wide variations in what criteria tutors thought were important and in how they actually marked the essays".

Preliminary attempts have also been made to measure different aspects of students' work, and to examine the relationships between those measures and overall marks. Newstead and Dennis (1994) asked 14 external examiners to rate six answers to the examination question "Is there a language module in the mind?" for 1. Quality of argument, 2. Extent, accuracy and relevance of knowledge displayed, 3. Level of understanding, 4. Insight, originality and critical evaluation, and 5. Relevance to and success in answering the question. In an analysis of variance there was no significant interaction between essays and aspects, "suggesting that markers did not have a common view of where the strengths and weakness of each script lie" (p. 218). In multiple regression with aspect ratings as the predictor variables and final marks as the outcome variable, all aspects except level of understanding were significantly related to final mark.

Using a similar approach but focusing on different aspects, Norton et al. (1999) asked markers of level 3 coursework essays in psychology to rate each essay for the student's effort, ability and motivation. The correlation between those ratings and the grade given to the essays was .81 for effort, .84 for ability and .79 for motivation. The three ratings were then used in multiple regression to predict grades awarded. Both ability and effort were significant independent predictors of grades. In another study, Norton (1990) asked students to complete a questionnaire about more factual aspects of the work they had done on coursework essays in psychology, and related those responses to essay grades. None of the factors that were examined (time spent, numbers of sources of material, numbers of drafts) were significantly related to marks awarded. A detailed analysis of the essays themselves, using a smaller sample of 10 high scoring and 10 low scoring essays did reveal differences, however. Number of references cited was significantly correlated with grade, and so were measures derived from a sentence-by-sentence content analysis of the essays, in which each sentence was assigned to one of 10 categories that were then collapsed to three: 'factual descriptive

information', 'research-based information', and 'structuring'. This produced scores representing the percentage of sentences in the essay assigned to each category. Factual descriptive information and research-based information were significantly related to essay grade, but structuring was not.

One obstacle to the use of cues or aspects such as those identified by Newstead and Dennis (1994) and Norton et al. (1999) to investigate examination marking is that for examinations there is almost never an external criterion against which to compare marks. For this reason most of the empirical research on marking has focused on the reliability rather than the validity of marks awarded. Where two or more markers make independent assessments of the same piece of work, psychometric methods can be used to estimate the 'precision' of the assessment or the extent of measurement error. Laming (1990) examined the marks awarded by pairs of markers for answers in a university examination over two years. The correlations between the two marks ranged from .47 to .72 for one year and from .13 to .37 for the second. Laming applied classical test theory to estimate the precision of the examination, and concluded that for the second year this was insufficient to support the published division of the class list.

Newstead and Dennis (1994) examined the reliability of the marks awarded by 14 external examiners and 17 internal markers to six answers to the examination question "Is there a language module in the mind?" The standard errors of measurement were 6.2 percentage points for the external examiners and 5.1 for the internal markers, and the coefficients of concordance were .46 and .58 respectively. Those levels of agreement were disturbing, but Newstead and Dennis argued that as students' degree classes are assessed over a number of examinations rather than just one, measurement error like that would be likely to lead to misclassification only for students who were very close to degree class borderlines.

That view was supported by Dracup's (1997) analysis of psychology degree marking. Combining the different components of assessment for each unit, the correlations between marks awarded by first and second markers ranged from .47 to .93 for compulsory units. They were much more variable, including several that were not significantly correlated, for optional units with smaller numbers of students. However, when the marks across all the units were averaged, the correlation between the averages of the first and second marks was .93, a much more encouraging level of agreement.

The level of agreement between markers, or the reliability of marking, says little about the validity of the marking, except that the validity cannot be greater than the reliability. Questions about the validity of marking are raised, for example, by evidence of differences in psychology degree classifications between institutions or between different years (eg., Myron-Wilson & Smith, 1998; Smith, 1990), and by evidence that marks may be affected by personal knowledge of the student (eg., Dennis et al., 1996). However, understanding the sources of differences between markers can go some way towards improving the quality of marking. Laming (1990, p. 247) observed that markers may remind themselves during marking of what they are looking for in answers, so that they employ notional model answers, and that markers with different areas of professional expertise would adopt different model answers as a basis for their judgements. In many cases the difference between two markers is that the 'first marker' is the person who taught the material being examined and set the question for students to answer, and the 'second marker' is a person who is more broadly familiar with the material being examined. In those situations the first marker might be expected to have much clearer expectations about how the question could be answered, and judgement analysis could provide insights into the ways in which the different perspectives of first and second markers affect the marks they award.

In the present study we extended the approaches of Newstead and Dennis (1994) and Norton et al. (1999) to conduct a more formal application of judgement analysis to the marking of examination answers in psychology. The study took advantage of the development of very detailed assessment criteria that specified levels of achievement for each of seven aspects of examination answers. These were 1. *Addresses the question,* 2. *Covers the area*, 3. *Understands the material,* 4. *Evaluates the material*, 5. *Presents and develops arguments*, 6. *Structures answer and organises material,* and 7. *Clarity in presentation and expression*. For each aspect, the criteria described standards for seven levels of achievement corresponding to grade bands (see appendix).

The criteria were consistent with published descriptions of good practice in student assessment (eg., Miller et al., 1998; Quellmalz, 1991) and previous research on essay writing and student assessment in psychology (eg., Norton, 1990; Norton et al., 1996a, 1996b, 1999), and were developed through discussion and consultation within the department (Elander, 2002). The aim was to identify a small number of aspects or attributes of students' coursework essays and examination answers that staff believed were important factors that they considered in awarding marks and that could potentially be assessed independently of one another. The criteria were by used markers to promote more reliable marking and were incorporated in course materials to support students' learning and promote 'deeper learning strategies' (Marton & Saljo, 1976) by setting out the qualities that markers look for in coursework and examination answers.

The specification of potentially independent aspects of assessment and their adoption at departmental level allowed several empirical issues to be investigated. Firstly, the extent to which markers are able to make ratings of those aspects that are statistically independent of one another can be examined. Secondly, judgement analysis can be used to 'capture the policies' of individual markers and identify reasons for differences between the marks awarded by first and second markers. Thirdly, a model consisting of a combination of specific aspect ratings can be tested by examining the relative contributions made by the model and overall marks in the prediction of an external criterion, such as the mark awarded by another marker.

The study involved actual examination answers over a range of university psychology examinations. The markers made separate ratings for the seven aspects of the assessment, as well as recording their overall mark for each answer. The overall mark awarded by the co-marker was also recorded. The aims of the study were as follows:

1. To examine the underlying structure of aspect ratings. We used principal components analysis to assess the extent to which variations in aspect ratings could be accounted for by a smaller number of components. This analysis focuses on the three tasks of the expert judge (as described by Einhorn, 2000) that precede the integration of information, which are to identify relevant cues, to measure the amount of the cues, and to cluster information about cues. The reason for this focus is that the cues, or aspects, specified in the assessment criteria cannot for the present be verified objectively in the same way as the cues employed in most of the research where expert judgement has been compared with statistical models.
2. To describe, or 'capture,' the judgement 'policies' of individual markers. In multiple regression analyses, we used aspect ratings to predict overall marks awarded for each marker. We wished to know two things in particular about each analysis. The first was how much of the variance in overall marks was accounted for (and conversely how much was unexplained) by the ratings of specific aspects of the assessment criteria. The second was how many aspects were independently associated with aspect ratings, and therefore how well the overall marks incorporated specific aspect ratings (aspects with significant independent associations with

overall marks being those that were reflected in the overall mark). We also wished to compare markers acting as first and second markers. First markers had taught the material being examined and set the question, so we expected them to be in a better position to award overall marks that reflected aspects of the assessment criteria. We predicted that the proportion of variance in overall marks accounted for by aspect ratings, and the number of aspect ratings independently associated with overall marks, would be greater for first markers than second markers. We made no prediction, however, about differences between first and second markers in the roles played by particular aspect ratings in accounting for overall marks.

3. To apply a simple ('improper') linear model consisting of the sum of the aspect ratings, and assess the extent to which that model added to overall marks in the prediction of marks awarded by a separate marker working independently (the 'co-marker'). We tested the increase in variance in co-markers' overall marks accounted for by the sum of the seven aspect ratings, over and above that accounted for by the overall marks of the person making aspect ratings. This was achieved by comparing the prediction of co-markers' marks by the sum of the aspect ratings as well as the overall marks with their prediction by the overall marks alone. Because we expected the overall marks of first markers to reflect aspect ratings to a greater extent, we predicted that the increase in variance accounted for by aspect ratings would be greater for those made by second markers.

Methods

Seven full-time members of the department's academic staff rated examination answers on the seven aspects of the assessment criteria, as well as providing an overall mark for each answer. The markers volunteered to take part in the exercise and the marking formed part of their usual examination marking workload. The data were collected during three examination sessions (Autumn and Spring semester papers, plus the Summer resit examinations) in a single academic year. All the assessment was conducted blind to the students' identities.

Details of the marking are given in table 1. The course units included eight from the undergraduate programme in psychology and two from an MSc course in Occupational Psychology. The examinations all required students to attempt three out of eight questions in two hours, providing answers in the form of short essays. There were 551 answers, with 258 assessed by markers acting as the 'first marker' (the member of staff who had taught the material and set the question), and 293 by markers acting as the 'second marker' (a member of staff with more general expertise in the area of the examination). For 322 answers, the overall mark awarded by a co-marker who did not make aspect ratings was available. In all of those cases, both markers conducted their marking blind to the marks awarded by the other.

The number of questions from each paper marked by staff making aspect ratings ranged from one to eight, and the number of answers from each paper ranged from 11 to 119. The markers were instructed to make judgements about the quality of answers in terms of the seven aspects of the assessment criteria at the same time that they arrived at an overall mark for each answer. They were asked not to attempt to change the way they arrived at overall marks, and were not asked to give equal weight to each of the aspects. Ratings were made on a seven-point scale (1=low, 7=high), with each point corresponding to a level of achievement specified in the criteria (see appendix). Overall marks were made on a percent scale.

Table 1. Marking details

Marker	Papers (level) marked	Number of questions marked	Number of answers marked
A	Decision Making and Choice (UG level 3)	8	86
	Decision Making and Choice - Resit (UG level 3)	7	11
B	Social Psychology (UG level 3)	4	36
C	Cognitive and Developmental Psychology (UG level 1)	8	69
	Memory and Understanding (UG level 2)	3	119
	Memory and Understanding - Resit (UG level 2)	1	14
	Thought and Language - Resit (UG level 3)	5	10
D	Cognitive Psychology (UG level 1)	8	45
E	Adult development and Ageing (UG level 3)	1	29
F	Ergonomics (PG)	5	42
	Social Research in Applied Settings (PG)	7	57
G	Developmental Psychology - Resit (UG level 2)	8	33

Note: UG = Undergraduate Psychology Programme, PG = MSc Occupational Psychology

Results

Four answers marked by second markers were excluded from the analysis because of missing data on one or more aspect rating. The data analysis was conducted with SPSS version 10.0

Table 2 shows the means and standard deviations for each marker's aspect ratings and overall marks. Separate figures are reported for answers where the marker acted as the first marker (marking the questions they had set themselves) and as the second marker (marking questions set by someone else). Mean aspect ratings ranged from 2.49 to 6.38, and standard deviations from 0.57 to 1.93. Mean overall marks ranged from 43.2 to 65.9, and standard deviations from 5.0 to 18.7. When data from all the markers were combined, mean aspect ratings ranged from 4.05 to 5.01, and standard deviations from 1.31 to 1.59. Among the four markers who marked both their own and others' questions (Markers A, C, F and G), there was a tendency to higher aspect ratings when acting as a second marker (25 out of 28 cases), and to show less variability (22 out of 28 cases). Overall marks also tended to be higher for second-marked questions (marginally so for Marker G).

There were also differences between markers. Overall marks and aspect ratings were higher for marker E, who marked just one well-answered question, and for marker F, who marked two masters level examinations. Markers E and F also showed much less variability than the other markers in their aspect ratings (all standard deviations were less than one) and their overall marks. Marker D, who acted only as a second marker, tended to give quite low aspect ratings, but the mean overall mark (50.8%) was by no means the lowest among the markers.

Underlying structure of aspect ratings

Principal components analysis was applied to the aspect ratings made by each marker and all the markers combined. Table 3 shows the eigen values and percent of the total variance accounted for by each of seven potential components. This showed that in every case the first factor accounted for the majority of total variance in ratings (the figures ranged from 56.5% for marker F to 80.6% for marker A), but that the eigen values of second components was very close to 1.0 for two markers (C and G). This would indicate that a single component accounted for most of the variation in aspect ratings. The scree plots broadly confirmed the importance of first components in each analysis (figure 1). In several cases, however, the scree plots indicated a clear, if minor, role played by second or third components, notably for markers C and G. This was true to a lesser extent for marker B and for all the markers combined. It should be noted, moreover, that principal components analysis is designed to identify first components that maximise the proportion of total variance accounted for, and would be expected to underestimate the importance of subsequent factors (Kline, 1994).

Table 4 shows the loadings of the seven aspect ratings on the first three components extracted. The highest loadings for each component in each analysis are shown in bold. Loadings on the first component were highest for aspects 1, 2, 3, 4 and 5. Loadings on the second component were highest for aspects 6 and 7, with the exceptions of markers E and G. For both of those markers, however, aspects 6 and 7 loaded highly on component three. Loadings on the third component were more mixed, apart from marker G, and to a lesser extent markers C and E, for whom the third component resembled the other markers' second component.

The principal components analysis did not provide a strong case for more than one clear component, especially when the data from all markers were combined. However, markers varied considerably in the component structure underlying their ratings, and this may indicate differences in their ability to make independent assessments of different aspects of the examination essays. The clearest pattern was produced by marker G, whose ratings justified a three component structure, with aspects 3, 4 and 5 loading on one component, aspects 1 and 2 on a second, and aspects 6 and 7 on a third.

Capturing the judgement policies of markers

Multiple regression, with aspect ratings as predictor variables and overall marks as outcome variables, was used to examine the relative influence of the seven aspects on marks awarded, using standard judgement analysis methods. The results should be treated with a certain amount of caution in the light of the results of the principal components analysis, which reflect correlations among the aspect ratings. The correlations among aspect ratings for individual markers ranged from .29 to 89, and for all of the markers combined they ranged from .59 to .86. The correlations between aspect ratings and overall marks for individual markers ranged from .45 to .96 for first markers and from .39 to .93 for second markers, and for all markers combined ranged from .65 to .91 for first markers and from .59 to .82 for second markers. We present the results as an illustration of the way that quantitative measures derived from assessment criteria can be used to investigate markers' judgements, and as the basis for hypotheses about those judgements that can be tested in further analyses.

Table 2. Means and standard deviations (in italics) of aspect ratings and overall marks. The first figures are for answers where the marker acted as the first marker (question-setter), those in parenthesis where they acted as second marker.

Aspects of the Assessment Criteria	Marker A N = 33 (64)	Marker B N = 36	Marker C N = 122 (90)	Marker D N = (45)	Marker E N = 29	Marker F N = 25 (74)	Marker G N = 13 (20)	All markers N = 258 (293)
1. Addresses the question	4.27 (4.44) *1.77 (1.42)*	5.00 - *1.47 -*	4.53 (5.28) *1.52 (1.06)*	- (4.31) *- (1.62)*	6.38 - *0.62 -*	5.28 (5.81) *0.84 (0.82)*	4.00 (3.95) *1.53 (1.91)*	4.82 (5.00) *1.55 (1.41)*
2. Covers the area	3.91 (4.41) *1.86 (1.58)*	3.91 - *1.58 -*	4.29 (4.34) *1.51 (1.22)*	- (3.47) *- (1.47)*	5.97 - *0.87 -*	4.80 (5.23) *0.91 (0.80)*	3.38 (3.50) *1.39 (1.93)*	4.39 (4.39) *1.59 (1.45)*
3. Understands the material	4.12 (4.58) *1.78 (1.55)*	4.09 - *1.38 -*	4.16 (4.67) *1.37 (0.97)*	- (3.53) *- (1.65)*	5.66 - 0.67 -	5.00 (5.19) *0.76 (0.77)*	3.46 (4.25) *1.39 (1.25)*	4.36 (4.58) *1.43 (1.32)*
4. Evaluates the material	3.91 (3.72) *1.79 (1.46)*	4.18 - *1.49 -*	3.79 (4.17) *1.29 (1.13)*	- (2.49) *- (1.22)*	5.45 - *0.91 -*	4.80 (5.12) *0.65 (0.59)*	3.31 (4.15) *1.60 (1.23)*	4.12 (4.05) *1.44 (1.40)*
5. Develops arguments	3.42 (4.08) *1.50 (1.36)*	4.03 - *1.31 -*	4.26 (4.49) *1.37 (1.19)*	- (2.67) *- (1.13)*	5.66 - *0.81 -*	4.72 (4.99) *0.61 (0.61)*	3.31 (4.00) *1.49 (1.34)*	4.28 (4.21) *1.41 (1.34)*
6. Structures and organises material	4.52 (4.81) *1.77 (1.45)*	4.51 - *1.44 -*	4.51 (4.77) *1.44 (1.21)*	- (2.73) *- (1.39)*	5.79 - 0.77 -	4.80 (5.24) *0.71 (0.57)*	3.54 (4.20) *1.27 (1.77)*	4.63 (4.55) *1.31 (1.46)*
7. Clarity in presentation	4.69 (5.03) *1.70 (1.41)*	4.49 - *1.36 -*	5.14 (5.14) *1.19 (0.95)*	- (3.00) *- (1.37)*	6.10 - 0.90 -	4.92 (5.32) *0.76 (0.58)*	3.58 (4.25) *1.31 (1.59)*	5.01 (4.78) *1.34 (1.37)*
Overall mark	47.7 (51.1) *17.9 (18.7)*	46.5 - *16.7 -*	49.8 (53.9) *17.0 (11.9)*	- (50.8) *- (14.9)*	65.9 - *5.3 -*	53.0 (56.7) *5.9 (5.0)*	43.2 (43.9) *14.0 (12.9)*	50.9 (52.8) *16.2 (13.4)*

Table 3. Eigen values and percent of variance accounted for (shown in brackets) for seven components extracted by principal components analysis of aspect ratings

	Marker A	Marker B	Marker C	Marker D	Marker E	Marker F	Marker G	All Markers
Component								
1	5.6 (80.6)	5.5. (78.6)	4.8 (68.8)	6.15 (87.9)	4.8 (69.4)	4.0 (56.5)	4.5 (63.9)	5.3 (76.3)
2	.41 (5.8)	.63 (8.9)	.998 (14.3)	.28 (4.0)	.68 (9.7)	.86 (12.3)	.94 (13.4)	.63 (9.0)
3	.28 (4.1)	.29 (4.1)	.36 (5.1)	.22 (3.2)	.50 (7.1)	.65 (9.2)	.93 (13.2)	.31 (4.4)
4	.25 (3.5)	.19 (2.8)	.30 (4.3)	.14 (1.9)	.36 (5.2)	.54 (7.8)	.24 (3.5)	.22 (3.2)
5	.18 (2.6)	.18 (2.6)	.21 (3.0)	.09 (1.3)	.29 (4.2)	.39 (5.6)	.18 (2.6)	.20 (2.8)
6	.14 (2.0)	.13 (1.8)	.18 (2.5)	.07 (1.0)	.20 (2.8)	.33 (4.7)	.15 (2.2)	.17 (2.4)
7	.09 (1.3)	.08 (1.2)	.14 (2.0)	.05 (0.7)	.11 (1.6)	.28 (4.0)	.09 (1.3)	.14 (2.0)

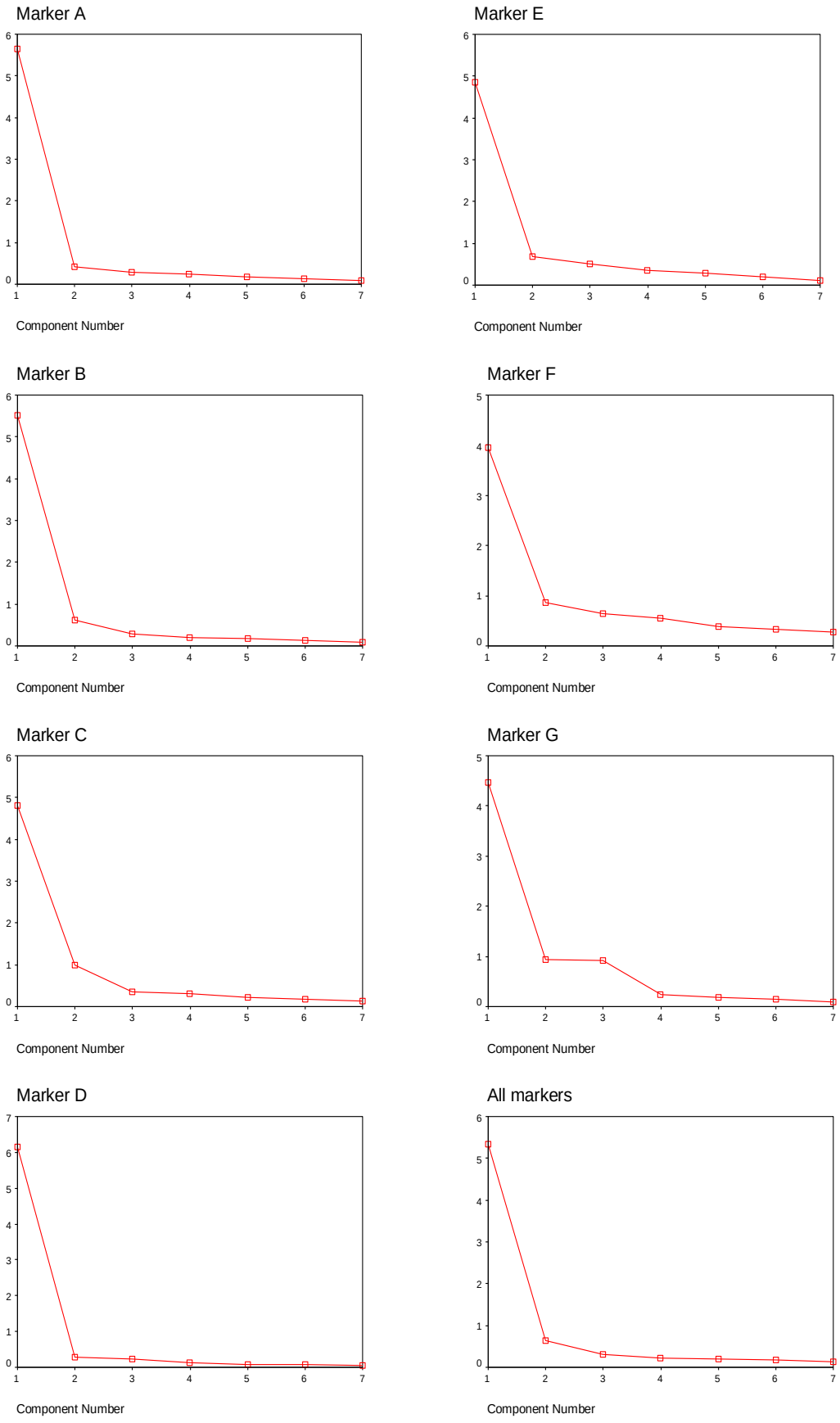

Figure 1. Scree plots showing eigen values for seven components extracted from principal components analysis of aspect ratings for markers A – G and all markers combined.

Table 4. Loadings of aspect ratings 1-7 on three components extracted from principal components analysis for markers A-G and all markers combined.

	Marker A	Marker B	Marker C	Marker D	Marker E	Marker F	Marker G	All Markers
Aspect rating				Component 1				
1	.81	**.93**	**.87**	**.93**	**.89**	**.90**	.43	.74
2	**.93**	**.92**	**.91**	**.97**	**.91**	**.87**	.59	**.87**
3	**.92**	**.93**	**.93**	**.98**	**.86**	**.88**	**.90**	**.89**
4	**.93**	**.92**	**.91**	**.97**	.76	.48	**.93**	**.95**
5	**.93**	**.91**	**.93**	**.91**	.59	.59	**.93**	**.94**
6	.80	.64	.56	.86	.70	.55	.54	.72
7	.76	.75	.42	.84	.72	.39	.51	.63
				Component 2				
1	.75	.54	.48	.88	.20	.48	**.97**	.62
2	.78	.74	.22	.87	.06	.49	**.89**	.67
3	.82	.66	.32	.84	-.24	.45	.66	.69
4	.70	.60	.31	.84	**.57**	.44	.36	.66
5	.77	.68	.31	.87	.12	.64	.40	.69
6	**.96**	**.98**	.55	**.96**	-.39	**.89**	.33	**.94**
7	**.96**	**.87**	**.97**	**.97**	.03	**.85**	.46	**.96**
				Component 3				
1	**.76**	-.09	-.44	-.22	.66	.33	-.34	**.97**
2	.44	-.24	-.52	.08	.58	.37	-.68	.86
3	.45	-.29	-.53	.12	.69	.48	-.53	.85
4	.43	-.39	-.45	.22	.67	**.96**	-.54	.71
5	.13	-.12	-.50	**.38**	**.94**	.59	-.47	.69
6	.22	-.07	**-.99**	.17	**.86**	.25	**-.96**	.58
7	.35	**-.53**	-.58	-.10	**.89**	.48	**-.95**	.57

Tables 5 and 6 show the results of a series of stepwise multiple regression analyses in which aspect ratings were used to predict overall marks for markers acting as first markers (table 5) and second markers (table 6). The criterion for entry into a regression model was $p < .05$, and the criterion for removal was $p > .1$. In each case we report the adjusted R^2, as is appropriate when comparing across regression equations involving different sample sizes and different numbers of independent variables (Hair et al., 1998, p. 182). Hair et al. also give the values of R^2 that can be detected as a function of sample size, significance level, and number of independent variables. The smallest sample size they consider is 20. At $\alpha = .05$, and with 5 independent variables, $R^2 \geq .48$ will be detected 80% of the time, and with 10 independent variables, $R^2 \geq .64$ will be detected 80% of the time (Hair et al., 1998, p.165).

Table 5. Aspect ratings that made significant independent contributions to the prediction of overall marks for answers where the marker acted as the first marker (question-setter). The table shows standardised regression coefficients (β) from final models for aspect ratings and proportions of variance in overall marks accounted for (R^2) in stepwise multiple regression models in which aspect ratings were used to predict overall marks. 'M' shows the regression models in which the aspect ratings made a significant independent contribution.

Aspects of the criteria	Marker A (N = 33)	Marker B (N = 33)	Marker C (N = 122)	Marker E (N = 29)	Marker F (N = 25)	Marker G (N = 12)	All first markers (N = 254)
1. Addresses the question	β = .240** M = 2, 3		β = .190*** M = 3, 4, 5		β = .260** M = 5, 6		β = .168*** M = 3, 4, 5, 6
2. Covers the area	β = .232* M = 3	β = .292** M = 1, 2, 3, 4	β = .405*** M = 2, 3, 4, 5	β = .172* M = 5	β = .254** M = 2, 3, 4, 5	β = .681*** M = 1, 2	β = .342*** M = 1, 2, 3, 4, 5, 6
3. Understands the material	β = .543*** M = 1, 2, 3		β = .190*** M = 1, 2, 3, 4, 5	β = .215* M = 2, 3, 4, 5	β = .157[a] M = 1, 2, 3, 4, 5, 6		β = .194*** M = 2, 3, 4, 5, 6
4. Evaluates the material		β = .350*** M = 3, 4		β = .305*** M = 3, 4, 5			β = .089*** M = 6
5. Develops arguments			β = .204*** M = 4, 5		β = .208*** M = 3, 4, 5, 6	β = .396** M = 2	β = .161*** M = 4, 5, 6
6. Structures and organises material		β = .190* M = 4	β = .090*** M = 5	β = .233* M = 4, 5	β = .125* M = 6		β = .097*** M = 5, 6
7. Clarity in presentation		β = .251** M = 2, 3, 4		β = .232* M = 1, 2, 3, 4, 5	β = .231*** M = 4, 5, 6		
Model 1: R^2	.920	.814	.841	.718	.730	.845	.822
Model 2: R^2	.933	.900	.902	.860	.873	.946	.878
Model 3: R^2	.941	.923	.925	.902	.935		.895
Model 4: R^2		.935	.933	.917	.945		.904
Model 5: R^2			.938	.930	.959		.908
Model 6: R^2					.967		.910

a $p = .061$ * $p < .05$ ** $p \leq .01$ *** $p \leq .001$

Table 6. Aspect ratings that made significant independent contributions to the prediction of overall marks for answers where the marker acted as the second marker. The table shows standardised regression coefficients (β) from final models for aspect ratings and proportions of variance in overall marks accounted for (R^2) in stepwise multiple regression models in which aspect ratings were used to predict overall marks. 'M' shows the regression models in which the aspect ratings made a significant independent contribution.

Aspects of the assessment criteria	Marker A (N = 64)	Marker C (N = 90)	Marker D (N = 45)	Marker F (N = 74)	Marker G (N = 20)	All second markers (N = 293)
1. Addresses the question				β = .239*** M = 6, 7		β = .268*** M = 2, 3
2. Covers the area	β = .604*** M = 1, 2	β = .262*** M = 1, 2, 3, 4	β = .683*** M = 1, 2	β = .260*** M = 3, 4, 5, 6, 7	β = .672*** M = 1, 2	β = .407*** M = 1, 2, 3
3. Understands the material		β = .385*** M = 2, 3, 4		β = .195*** M = 1, 2, 3, 4, 5, 6, 7		β = .230*** M = 3
4. Evaluates the material		β = .306*** M = 3, 4		β = .192*** M = 5, 6, 7		
5. Develops arguments				β = .178*** M = 2, 3, 4, 5, 6, 7	β = .393** M = 2	
6. Structures and organises material				β = .106** M = 7		
7. Clarity in presentation	β = .262* M = 2	β = .138*** M = 4	β = .293** M = 2	β = .161*** M = 4, 5, 6, 7		
Model 1: R^2	.634	.764	.853	.613	.704	.666
Model 2: R^2	.659	.826	.878	.749	.824	.701
Model 3: R^2		.863		.817		.714
Model 4: R^2		.879		.871		
Model 5: R^2				.900		
Model 6: R^2				.927		
Model 7: R^2				.934		

* $p < .05$ ** $p \leq .01$ *** $p \leq .001$

Because of the correlations among aspect ratings, we obtained collinearity diagnostics (Belsley et al., 1980) for each analysis. These give the 'tolerance' (1 minus the squared multiple correlation for each variable with the rest as predictor variables in multiple correlation, so that low tolerances indicate high collinearity), and a conditioning index and variance proportions associated with each variable, after standardisation, for each root. The criteria for multicollinearity causing statistical instability are a conditioning index greater than 30 and more than two variance proportions greater than .5 for a given root number (Tabachnick & Fidell, 1996, p. 87).

Those data showed that the criteria for multicollinearity were met by only two of the regression analyses. Those were the analyses for markers E and F as first markers (table 5). For marker E the lowest tolerance in the final model was .285, and there was a conditioning index of 49 with associated variance proportions of .68 and .78. For marker F the lowest tolerance in the final model was .177, and there was a conditioning index of 56 with associated variance proportions of .67 and .84. For all of the other 11 analyses reported in tables 5 and 6, multicollinearity was acceptable, and the lowest tolerances in the final models ranged from .113 to .80, well above the highest default tolerance level (.01) employed by statistical programmes (Tabachnick & Fidell, 1996, p. 86).

The first thing to note about the results themselves is that aspect ratings accounted for substantial proportions of the variance in overall marks, especially for individuals acting as first markers, where the lowest R^2 for a final model was .93. For those individuals who undertook both first and second marking (markers A, C, F & G), there was much more unaccounted-for variation in overall marks for the second marking, where the lowest R^2 for a final model was .66. Space considerations mean that we have only shown the beta values for the final models. In each case, this is considerably less than in the initial model, reflecting inter-correlation among aspect ratings. The most dramatic example of this is for marker F acting as a first marker (table 5), whose first regression model involved aspect 3 with a beta value of .861. By the time the sixth model had been constructed this value had fallen to .157. For each marker the first regression model accounted for a considerable amount of the variance, with relatively modest increases in subsequent models. Table 5 also indicates considerable differences between markers, both in terms of the number of predictive aspects and the extent of their predictiveness. Marker A seemed to place most emphasis on *understands the material* whereas markers C and G placed most emphasis on *covers the area*. For other markers, there did not appear to be any aspects that were singularly important. *Covers the area* was the only aspect to be identified as a predictor of each marker's overall marks. *Evaluates the material* was the aspect that appeared in the fewest regression models (2 out of the 6).

In the second-marker data (table 6), *addresses the question* and *structures the material* entered into only one marker's regression model (marker F). Also, *understands the material* entered into just two markers' models, compared with 4 models in the first-marker data. There were also some differences in emphasis. For marker A, *understands the material* had a beta value of .543 in the first-marker data, whereas in the second-marker data not only did this aspect not enter the model, but *covers the area* had a beta value of .604. Different regression models were not always produced when markers acted as first and second markers, however. For marker G, overall marks were predicted by *covers the area* and *develops arguments* in both cases. Marker F's regression models included more aspect ratings (six as a first marker and seven as a second marker) and accounted for more of the variance in overall marks (.97 as a first marker and .93 as a second marker) than for any other marker. In both the first-marker and second-marker data, *covers the area* was the only aspect identified as an independent predictor of overall marks for every marker.

Tables 5 and 6 also show the results of regression analyses for data combined across all the markers. These showed that for first markers every aspect of the criteria except *clarity in presentation* made a significant independent contribution to the prediction of overall marks, whereas for second markers only *addresses the question, covers the area,* and *understands the material* were independently predictive of overall marks. In both cases*, covers the area* was the first aspect rating to be entered in the regression models, and had the highest beta

value in the final model. As with the results for individual markers, aspect ratings accounted for more of the variance in overall marks for first markers than for second markers, with R^2 values of .91 compared with .71.

Those analyses showed that judgement policies varied from marker to marker, but the differences between first and second markers were consistent with the expectation that first markers were more able than second markers to award overall marks reflecting the range of aspects specified in the assessment criteria. One limitation is that the analyses all use the markers' own overall mark as the variable to be predicted by their aspect ratings. If markers produced aspect ratings that reflected their overall marks rather than being independent of them, this might be reflected in the results. In the next analyses, we used aspect ratings made by one marker to predict overall marks awarded by another marker, so that aspect ratings and overall marks were made completely independently of one another. Another potential limitation is that the aspect ratings were correlated with one another. In the next analyses, we combined aspect ratings in a single summary score, so that correlated ratings were not used as separate predictor variables.

Testing an 'improper linear model' of marking judgement

For 322 of the answers, a second overall mark was available from a co-marker who did not make aspect ratings. In 143 of those cases, aspect ratings were provided by the first marker (the co-marker was the second marker), and in 179, aspect ratings were provided by the second marker (the co-marker was the first marker). The approach that we adopted was to test the contribution that an 'improper linear model,' consisting of the unweighted sum of the aspect ratings, made to the prediction of co-markers' overall marks. This was achieved by conducting regression analyses in two steps. In the first step, overall marks alone were regressed on co-markers' overall marks. In the second step, the sum of aspect ratings was added as a predictor variable and the increase in the amount of variance accounted for in co-markers' overall marks was tested. The analyses were then repeated to test the increase in variance accounted for by adding overall marks as a predictor to the sum of aspect ratings.

Because of the smaller numbers of answers for which there were two overall marks, analyses were not conducted separately for each marker. Instead, the data were combined across markers but analysed separately for cases where aspect ratings and overall marks by first markers were used to predict second markers' overall marks, and those where aspect ratings and overall marks by second markers were used to predict first markers' overall marks.

We again obtained collinearity diagnostics to guard against the results being distorted by correlations among the variables, applying the same criteria as before. These showed that collinearity was acceptable for all of the analyses. The tolerances in the second steps of each model were .092 where first marker data was used to predict the second marker, and .275 where second marker data was used to predict the first marker, with conditioning indices of 28 and 19 respectively.

Table 7 shows that for markers acting as first markers, the sum of the aspect ratings added almost nothing to the extent to which overall marks predicted co-markers overall marks. For markers acting as second markers, however, overall marks were less predictive of co-marks' overall marks, and including the sum of the aspect ratings in the regression equation added significantly to the prediction of co-makers' overall marks.

The analysis was repeated to assess the extent to which overall marks added to the sum of aspect ratings in the prediction of co-markers' overall marks (reversing the order in which overall marks and sum of aspect ratings were used as predictors in the two steps of the analysis). Table 8 shows that overall marks added significantly to the sum of aspect ratings in the prediction of co-markers' overall marks for both first and second markers, but to a much greater extent for first markers.

Those analyses show that the relative power of aspect ratings and overall marks to predict co-markers' overall marks differed between first and second markers. The increases in the proportions of variance in co-markers' marks accounted for (change in R^2) are plotted in fig. 2. The additional contribution of aspect ratings to

the prediction of marks awarded by co-markers was much greater when marks and aspect ratings by second markers were used to predict first markers. By contrast, the additional contribution of overall marks was greater when marks and aspect ratings by first markers were used to predict second markers. The results appear to indicate that for second markers, separate ratings of specific aspects of answers that were not incorporated in the overall mark awarded could help to explain discrepancies in marks between markers.

Table 7. Increase in the amount of variance in co-markers' overall marks accounted for by using the sum of aspect ratings in addition to overall marks as predictors.

	First marker data used to predict second marker (n = 139)			Second marker data used to predict first marker (n = 179)		
Predictor variables	R^2	Change in R^2	Sig. R^2 change	R^2	Change in R^2	Sig. R^2 change
Step 1. Overall mark	.710	.710	<.001	.661	.661	<.001
Step 2. Overall mark and sum of aspect ratings.	.710	.000	.950	.798	.137	<.001

Table 8. Increase in the amount of variance in co-markers' overall marks accounted for by using markers' overall marks in addition to the sum of markers' aspect ratings as predictors.

	First marker data used to predict second marker (n = 139)			Second marker data used to predict first marker (n = 179)		
Predictor variables	R^2	Change in R^2	Sig. R^2 change	R^2	Change in R^2	Sig. R^2 change
Step 1. Sum of aspect ratings.	.643	.643	<.001	.786	.786	<.001
Step 2. Sum of aspect ratings and overall mark.	.710	.067	<.001	.798	.012	.001

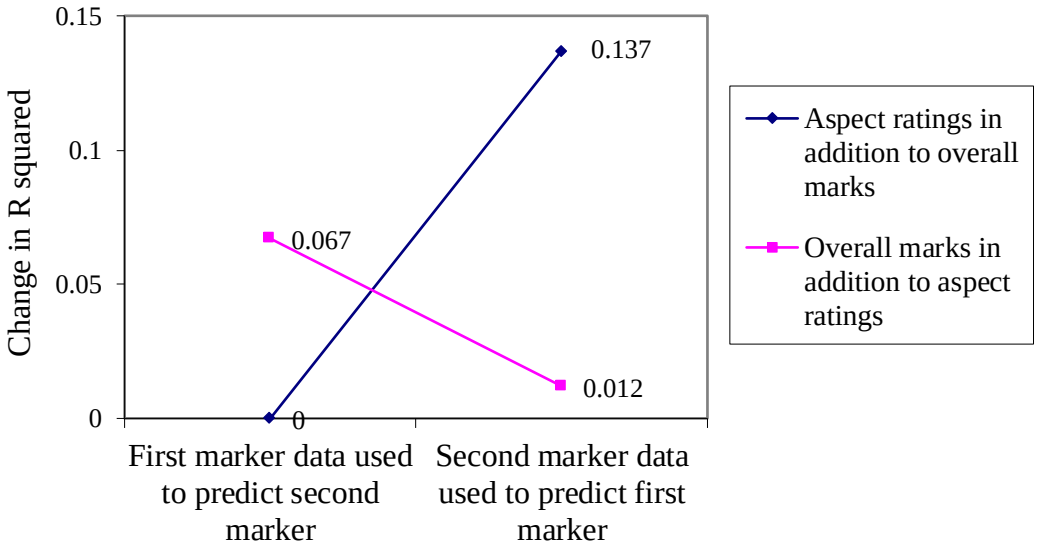

Figure 2. Increase in proportion of variance in co-markers' marks accounted for (change in R^2) by aspect ratings and overall marks compared with overall marks, and by aspect ratings and overall marks compared with aspect ratings.

Discussion

The results have implications for understanding the psychology of marking judgements and for the development of assessment criteria for educational purposes, and demonstrate the utility of using assessment criteria to generate detailed data for research on marking. The methods used are not without limitations, and the policy capturing analyses are illustrative rather than conclusive, but the use of an independent criterion and a much simpler set of predictors mean that much greater confidence can be placed in the results of the model testing analyses.

Psychometric status of aspect ratings

The principal components analysis did not provide clear evidence that markers were able to make ratings of separate aspects of answers that were statistically independent of one another. Principal components analysis, however, is an extremely conservative test of the extent to which ratings of the seven aspects were independent of one another. The method is designed to maximise the variance explained by the first component: "In most cases the first principal component explains far more variance than the other components. If most of the correlations in the matrix are positive, the first principal component has large positive loadings on most of the variables. This is called a general factor. That the first principal component is usually a general factor is an artifact of the method… It is thus inadmissible, but it is often done, to use the first principal component for the evidence of a general factor" (Kline, 1994, p. 39).

This part of the results is open to three interpretations about the validity of aspect ratings. The first is that they are not measures of seven different aspects, but little more than seven ratings of one common aspect, namely overall quality, and that this arose because markers were unable to make independent assessments of separate attributes. If this were the case, specifying separate aspects of assessment would not help to understand markers' judgements or improve the quality of marking (although there may still be educational benefits to presenting the criteria in this way, if they helped to remind students about important aspects of examination answers).

The second interpretation is that aspect ratings reflected the overall quality of answers not because markers were unable to make independent ratings but because they used aspect ratings to justify the overall mark awarded, or took their overall mark into consideration in other ways. Research in which markers made aspect ratings without determining an overall mark would determine whether this was the case, although in the third part of our analysis we were able to examine the relationships between aspect ratings made by one marker and overall marks awarded by another.

The third interpretation is that that markers could make independent aspect ratings but that the aspects were correlated in the students' work. There is no reason why answers that are given high ratings for one aspect should not also be given high ratings on another, even if the aspects are distinct and can be assessed independently of one another. The educational research that provided the basis for the aspects identifies them as conceptually distinct features that should be considered in judgements about overall marks (eg., Norton, 1990), but does not specify that they should not be correlated with each other, and all the evidence about the relationships between different aspects of students work shows that distinct aspects do tend to be correlated (Norton et al., 1999; Newstead & Dennis, 1994). In the one exception that we are aware of (Norton, 1990), the aspects were measured by counting the proportions of sentences in the essay that were assigned to mutually exclusive categories.

Further research will be needed to establish the validity of aspect ratings made by markers. As things stand, they could be said to have face validity (they describe constructs that are familiar and meaningful to the markers) and content validity (they are described in terms very close to those reported elsewhere, eg., Norton, 1990), but not criterion or construct validity, where an external criterion is required. That could be investigated by relating aspect ratings made by markers to a content analysis of the answers themselves, on lines similar to that employed by Norton (1990). The results might indicate acceptable validity only for a smaller number of aspects than were specified for the present study. The results of the principal components analysis indicated that three might be the upper limit, but that individual differences exist in the extent to which markers can differentiate aspects of assessment. Marker G, for example, provided the clearest three-component structure, where the components comprised aspects of deep learning (aspects 3, 4 & 5: *understanding*, *evaluation*, and *argumentation*), surface learning (aspects 1 & 2: *addressing the question* and *covering the area*), and presentation (aspects 6 & 7: *structure* and *clarity*). Those three broader aspects could form the basis for a simplified set of assessment criteria for future research on the validity and utility of assessment criteria. The implications of this are important, for they imply that in self-reports (eg., Norton, 1990), markers may overstate the number of separate attributes of essays they are able to consider in marking.

In addition to the number of aspects specified in the criteria, one might also question the number of points specified on the ratings scales. The seven levels employed here correspond to the five degree class bands, plus two levels of fail (compensatable and non-compensatable). While it is possible, administratively, to set out a complete matrix of criteria for all levels of all aspects, however, it is quite another for markers to use all of those levels in the appropriate way, and markers' use of the aspect ratings scales would require corroboration in further research. Early psychophysical research on judgements about amplitude, frequency and length of sensory stimuli showed that about five response categories were the most that judges could use without error in the absence of anchors (Laming, 1984). For subjective judgements, using scales with more points may increase reliability and validity. Preston & Colman (2000) compared ratings for aspects of the quality of stores and restaurants using scales with up to 11 response categories. Reliability and validity were significantly better with higher numbers of response categories, up to about seven, than with two-point, three-point, or four-point scales. Again, further research on the psychometric properties of aspect ratings will be needed to establish the optimal number of response categories that markers are able to use effectively.

Capturing the judgement policies of markers

The policy capturing analyses showed that the marks awarded by second markers were much less well predicted by their aspect ratings. Less of the variance in overall marks was accounted for, and fewer aspect ratings made significant independent contributions to the prediction of overall marks for second markers compared with first markers. Overall marks awarded by second markers therefore appeared to incorporate fewer separate aspects of the assessment, and depended more heavily on the aspect *covers the area.* This might be regarded as among the more superficial aspects of an answer and one that markers might reasonably be expected to consider before going on to consider whether answers had shown understanding, evaluation, argumentation and so on. First markers taught the material being examined and set the questions, and would be expected to be in a better position to award marks that reflected a wider range of attributes. They should have been better placed to make marking judgements that included *addressing the question* and *understanding the material*, and those aspect ratings were independently predictive of overall marks much more frequently for first markers than second markers. To some extent, therefore, those analyses provide tentative evidence of construct validity for aspect ratings, in that the results for first and second markers were consistent with expectations about the differing levels of expertise and familiarity with the material between first and second markers.

Multi-collinearity diagnostic statistics showed that for all but two of the policy capturing analyses, the degree of intercorrelation among predictor variables was below the level where the analysis would be compromised, but the policy capturing analyses should probably still be treated with a certain amount of caution. In most applications of judgement analysis the cues are independently verifiable, and our principal components analysis did not allow us to claim that aspect ratings were independent of one another. On the most conservative view, the policy capturing analyses illustrate how judgement analysis could be applied to examination marking using data generated by assessment criteria. They also provided the basis for a hypothesis that we were able to test in a much more rigorous way in the third part of the analysis.

Testing an improper linear model of marking judgement

In the policy capturing analyses, fewer aspect ratings made by second markers were independently associated with overall marks, so that second markers appeared to incorporate fewer aspects in their overall marks than did first markers. We therefore predicted that aspect ratings made by second markers would add more to the prediction of co-markers than those made by second markers. The two markers made their assessments without knowledge of one another's marks, so that co-markers' overall marks constituted an independent external criterion, and by using a model comprising the simple sum of aspect ratings we avoided the problem of using correlated aspect ratings as separate predictors. Using data from one marker to predict another addresses the issue of reliability between markers, and the analysis is an approach to explaining how discrepancies between markers arise.

The results supported the prediction, and, as fig 2 shows, the additional contribution made by aspect ratings to the prediction of co-markers' marks was almost zero for first markers but highly significant for second markers. Aspect ratings made by second markers, then, explained a significant portion of the variance in co-markers' marks that was unexplained by second markers' overall marks. Second markers were able to make ratings of specific aspects of answers that helped to predict first markers' marks but were not reflected in the marks they themselves awarded. The data support the argument that for second markers, ratings of specific aspects of examination answers would provide a more reliable measure of quality than an overall judgement. Indeed, for second markers, the sum of the aspect ratings accounted for more of the variance in co-markers' marks than did overall marks (tables 7 & 8 show R^2 values of .786 compared with .661). This was not the case for first markers, where the prediction of co-markers was not significantly improved by including aspect ratings as a

predictor, presumably because overall marks awarded by first markers incorporated aspect ratings to a much greater extent than for second markers.

Conclusions

These data provide preliminary evidence that measures of specific aspects of examination answers, appropriately combined, could be used to improve the reliability of marking, and provide an illustration of the ways that judgement analysis can be used to investigate the psychology of marking judgements. First and second markers appeared to differ in the extent to which the marks they awarded reflected specific aspects of assessment, and aspect ratings added significantly to the prediction of marks awarded by an independent marker, but only for second markers. If that pattern of results were supported by further research it would mean that the potential for improving reliability by calculating examination marks based on specific measures of performance may be limited to second markers. This would be consistent with the findings on expert judgement in other areas; Einhorn's classic research on clinical diagnosis, for example, showed that the predictive value of global judgements differed from judge to judge (Einhorn, 1972). Indeed, findings that point to marking procedures that would be differentially beneficial for first and second markers may usefully inform discussion about the administration of marking and the cost-effectiveness of double-marking (eg., Partington, 1994).

The findings are in line with those in many other areas where expert judgement has been examined, but a number of important cautions should be borne in mind. The aspects of assessment that were used in the present study require substantial further work. The underlying structure, reliability and validity of specific components of assessment all need to be established more fully. It may well be that a smaller number of aspects defined in somewhat different ways with different response scales will turn out to be a sounder basis for assessment than the seven aspects considered here. One way in which examination marking differs from almost all of the types of judgement where statistical combinations of specific measures were superior to global judgements is that for specific aspects of assessment as well as overall marks there is no clear external criterion or gold standard. This is one of the reasons why most empirical research on marking has been limited to the investigation of reliability rather than validity, and why the present results can speak directly only to the issue of reliability. The identification of validated aspects of assessment, especially those with lower inter-correlations, would allow research to test more sophisticated ways of combining predictor variables and investigate the validity of marking.

Acknowledgements

Many thanks to Liz Charman and everyone in the department of psychology for the design and development of the assessment criteria; to Robin Iwanek for suggesting that the psychology of expert judgement could be applied to marking; to the markers for their participation in the study; to the referees for their helpful comments on a previous draft of the paper; and to Nick Troop and Robert Parkinson for advice about the data analysis.

References

Belsley, D.A., Kuh, E. & Welsch, R.E. (1980). *Regression Diagnostics: Indentifying Influential Data and Sources of Collinearity.* New York: Wiley.

Carroll, J.S., Winer, R.L., Coates, D., Galegher, J. & Alibrio, J.J. (1988). Evaluation, diagnosis and prediction in parole decision making. *Law and Society Review,* 17, 199-228.

Cooksey, R.W. (1996). *Judgement Analysis: Theory, Methods and Applications*. London: Academic Press.

Dawes, R.M. (1994). *House of Cards: Psychology and Psychotherapy Built on Myth.* London: The Free Press.

Dawes, R.M. (1982). The robust beauty of improper linear models in decision making. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgement Under Uncertainty: Heuristics and Biases* (pp 331-407). Cambridge: Cambridge University Press.

Deacon, E.B. (1972). A discriminant analysis of prediction of business failure. *Journal of Accounting Research,* 10, 167-179.

Dennis, I., Newstead, S.E. & Wright, D.E. (1996). A new approach to exploring biases in educational assessment. *British Journal of Psychology,* 87, 515-534.

Dracup, C. (1997). The reliability of marking on a psychology degree. *British Journal of Psychology,* 88, 691-708.

Elander, J. (2002). Developing aspect-specific assessment criteria for examinations and coursework essays in psychology. *Psychology Teaching Review*, 10, 31-51.

Einhorn, H.J. (1972). Expert judgement and mechanical combination. *Organizational Behaviour and Human Performance,* 7, 86-106.

Einhorn, H.J. (2000). Expert judgement: some necessary conditions and an example. In T. Connolly, H.R. Arkes & K.R. Hammond (Eds.), *Judgement and Decision Making: An Interdisciplinary Reader* (2nd edition) (pp 324-335). Cambridge: Cambridge University Press.

Goldberg, L.R. (1968). Simple models or simple processes? Some research on clinical judgements. *American Psychologist,* 23, 483-496.

Goldman, L., Cook, E.F., Brand, D.A., Lee, T.H., Rouan, G.W., Weisberg, M.C., Acampora, D.A., Stasiulewicz, C., Walshon, J., Terranova, G., Gottlieb, L., Kobernick, M., Goldstein-Wayne, B., Copen, D., Daley, K., Brandt, A.A., Jones, D., Mellors, J. & Jakubowski, R. (1988). A computer program to prevent myocardial infarction in emergency department patients with chest pain. *New England Journal of Medicine,* 318, 797-802.

Hair, J.F., Anderson, R.E., Tatham, R.L., Black, W.C. (1998). *Multivariate Data Analysis* (5th edition). New Jersey, US: Prentice Hall.

Hoffman, P.J. (1960). The paramorphic representation of clinical judgement. *Psychological Bulletin*, 57, 116-131.

Kline, P. (1994). *An Easy Guide to Factor Analysis.* London: Routledge.

Laming, D. (1984). The relativity of 'absolute' judgements. *British Journal of Mathematical and Statistical Psychology,* 37, 152-183.

Laming, D. (1990). The reliability of a certain university examination compared with the precision of absolute judgements. *Quarterly Journal of Experimental Psychology,* 42A, 239-254.

Leli, D.A. & Filskow, S.B. (1981). Clinical-actuarial detection and description of brain impairment with the W-B Form 1. *Journal of Clinical Psychology*, 37, 623-629.

Marton, F. & Saljo, R. (1976). On qualitative differences in learning. 1. Outcome and process. *British Journal of Educational Psychology,* 46, 4-11.

Miller, A.H., Imrie, B.W. & Cox, K. (1998*). Student Assessment in Higher Education: A Handbook for Assessing Performance*. London: Cogan Page.

Myron-Wilson, R. & Smith, P.K. (1998). A matter of degrees. *The Psychologist*, 11, 535-538.

Newstead, S.E. & Dennis, I. (1994). Examiners examined: the reliability of exam marking in psychology. *The Psychologist,* 7, 216-219.

Norton, L.S. (1990). Essay writing: What really counts? *Higher Education*, 20, 411-442.

Norton, L., Brunas-Wagstaff, J. & Lockley, S. (1999) Learning outcomes in the traditional coursework essay: Do students and tutors agree? In C. Rust (Ed), *Improving Student Learning. Improving Student Learning Outcomes* (pp. 240-248). Oxford: The Oxford Centre for Staff and Learning Development.

Norton, L.S., Dickins, T.E. and McLaughlin Cook, A.N. (1996a). Coursework assessment: what are tutors really looking for? In G. Gibbs (Ed.), *Improving Student Learning: Using Research to Improve Student Learning* (pp. 155-166). Oxford: The Oxford Centre for Staff Development.

Norton, L.S., Dickins, T.E. and McLaughlin Cook, A.N. (1996b). Rules of the Game in essay writing. *Psychology Teaching Review,* 5, 1 -14.

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

Preston, C.C. & Colman, A.M. (2000). Optimal number of response categories in rating scale: reliability, validity, discriminating power and respondent preferences. *Acta Psychologica,* 104, 1-15.

Quellmalz, E.S. (1991). Developing criteria for performance assessments: the missing link. *Applied Measurement in Education,* 4, 319-331.

Smith, P.K. (1990). The distribution of psychology degree classes in the UK. *The Psychologist, Bulletin of the British Psychological Society,* 3, 147-152.

Tabachnick, B.G. & Fidell, L.S. (1996). *Using Multivariate Statistics* (3rd Ed.). New York: Harper Collins.

deVaul, R.A., Jervey, F., Chappell, J.A., Carver, P., Short, B. & O'Keefe, S. (1957). Medical school performance of initially rejected students. *Journal of the American Medical Association,* 257, 47-51.

Appendix. The assessment criteria developed for use in marking examination answers and coursework essays in psychology.

CORE ASSESSMENT CRITERIA	**1st (70 - 100%)**	**2:1 (60 - 69%)**	**2:2 (50 - 59%)**	**3rd (41 - 49%)**	**Pass (38 - 40%)**	**Fail (25 - 37%)**	**Fail (0 - 24%)**
Addresses question asked	Directly, synthesising appropriate material and showing insight into the issues raised by the question	Directly, synthesising appropriate material to provide an answer to the question	Somewhat indirectly, by presenting relevant material and trying to 'link' it to the question; some synthesis of material	Partially, by presenting material to answer part of the question but not all of it; some synthesis of material	At a general level, by presenting material on the topic but not addressing the question; little synthesis of material	Presents some material which *could* be related to question, but the question is ignored; no synthesis of material	Not at all, answers a different question
Covers the area	Very well, providing a comprehensive account of the material based on extended reading - including current literature	Well, providing accurate accounts of relevant material - clear evidence of reading beyond lecture notes and core texts	Satisfactorily, but some errors and/or omissions in accounts of relevant material - largely based on lecture notes and core texts	Adequately, but some significant errors and/or omissions in accounts of relevant material - no evidence of reading beyond lecture notes and core texts	Superficially, with significant errors and/or omissions in accounts of relevant models, theories etc.- evidence of bare minimum of required reading	Sketchily, providing partial descriptions of some of the material, but insufficient overall - little evidence of reading lecture notes or core texts	Does not cover the material, presenting only own ideas and/ or irrelevant material- no evidence of reading lecture notes or core texts
Understanding of material	Depth of understanding of conceptual, theoretical and methodological issues	Good understanding across the breadth of the material and some depth	Good understanding of the core material and some depth	Reasonable understanding of core material, but no depth	A general understanding of the material at a basic level, but no depth	No clear understanding of core material and evident confusion	Basic misunderstanding of core material

Appendix continued

CORE ASSESSMENT CRITERIA	1st (70 - 100%)	2:1 (60 - 69%)	2:2 (50 - 59%)	3rd (41 - 49%)	Pass (38 - 40%)	Fail (25 - 37%)	Fail (0 - 24%)
Evaluates the material	Insightful critical evaluation of the material and elaboration of alternative perspectives and current controversies	Evaluation includes conceptual/ methodological critique and an appreciation of alternative perspectives and current controversies	Some critical evaluation of material and an awareness of alternative perspectives and current controversies	Limited critical evaluation of material	Shows awareness of a critical perspective, but does not elaborate or discuss it	No evaluation of material	No evaluation of material or inappropriate criticisms rendered
Presents and develops arguments	Originality in arguments developed; good use of theory and empirical evidence in debate to present a strong case	Develops own argument; uses theory and empirical evidence to debate the case	Develops mainly derivative arguments; presents supporting evidence	Presents some arguments with supporting evidence, but doesn't develop these arguments	Presents some arguments, but at a superficial level and makes little use of supporting evidence	Presents poor, unsupported arguments	Presents no arguments, or ones which are clearly erroneous
Structures answer and organises material	Clear structure, material organised around the question asked	Clear structure, material organised well	General structure clear, but organisation of material muddled in places	Overall structure unclear, but sufficient 'flow' to the material	Confusing structure, limited organisation of material	Answer unstructured and material disorganised	Material presented as unconnected points, use of note form
Shows clarity and coherence in presentation and expression	Clear expression of ideas and cogent argumentation	Material and arguments presented clearly and coherently	Some minor points of lack of clarity and coherence	Adequate for comprehension, but generally insufficient concern for clarity and coherence	Some material presented unclearly or arguments lacking coherence	Material presented unclearly and arguments not coherent	Material presented unclearly, incoherent/ incomprehensible