with those of judges from the education sector, with the aim of also including representatives from these further stakeholder groups in Awarding.

References

Andrich, D. (1978). Relationships between the Thurstone and Rasch approaches to item scaling. *Applied Psychological Measurement*, **2**, 449–460.

Black, B. (2008). *Using an adapted rank-ordering method to investigate January versus June awarding standards*. Paper presented at the Fourth Biennial EARLI/ Northumbria Assessment Conference, Berlin, Germany, August 2008.

Bramley, T. (2005). A rank-ordering method for equating tests by expert judgment. *Journal of Applied Measurement*, **6**, 2, 202–223.

Bramley, T. (2007). Paired comparison methods. In: P. Newton, J-A Baird, H. Goldstein, H. Patrick and P. Tymms (Eds.), *Techniques for monitoring the comparability of examination standards* (chapter 7). London: QCA.

Bramley, T., Gill, T. and Black, B. (2008). *Evaluating the rank-ordering method for standard maintaining*. Paper presented at the 34th Annual Conference of the International Association for Educational Assessment, Cambridge, UK, September 2008.

Laming, D. (2004). *Human judgment: the eye of the beholder*. London: Thomson.

Linacre, J.M. (2006). *FACETS [Computer program, version 3.60.0]*. www.winsteps.com

Thurstone, L.L. (1927a). Psychophysical analysis. *American Journal of Psychology*, **38**, 368–389. In: Thurstone, L.L. (1959). *The measurement of values* (chapter 2). Chicago, Illinois: University of Chicago Press.

Thurstone, L.L. (1927b). A law of comparative judgment. *Psychological Review*, **34**, 273–286. In: Thurstone, L.L. (1959). *The measurement of values* (chapter 3). Chicago, Illinois: University of Chicago Press.

Wikipedia (2008). *Law of comparative judgment*. http://en.wikipedia.org/wiki/ Law_of_comparative_judgment Accessed 11 July 2008.

ASSESSMENT JUDGEMENTS

Using 'thinking aloud' to investigate judgements about A-level standards: Does verbalising thoughts result in different decisions?

Dr Jackie Greatorex and Rita Nádas Research Division

This article is based on a paper presented at the British Educational Research Association Annual Conference, September 2008, Edinburgh.

Abstract

Background

The 'think aloud' method entails people verbalising their thoughts while they do tasks, resulting in 'verbal protocols'. The verbal protocols are analysed by researchers to identify the cognitive strategies and processes as well as the factors that affect decision making. Verbal protocols have been widely used to study decisions in educational assessment. The main methodological concern about using verbal protocols is whether thinking aloud compromises ecological validity (the authenticity of the thought processes) and thus the decision outcomes. Researchers have investigated to what extent verbalising affected the thinking processes under investigation in a variety of settings. Currently, the research literature generally is inconclusive; most results show just longer performance times and no alternative task outcome.

Previous research on *marking* collected decision outcomes from two conditions:

1. marking silently;
2. marking whilst thinking aloud.

The mark to re-mark differences were the same in the two conditions. However, it is important to confirm whether verbalising affects decisions about grading standards. Therefore, our main aim was to compare the outcomes of senior examiners making decisions about *grading* standards silently as opposed to whilst thinking aloud. Our article draws from a wider project taking three approaches to grading.

Method

In experimental conditions senior examiners made decisions about A-level grading standards for a science examination both silently and whilst thinking aloud. Three approaches to grading were used in the experiment. All scripts included in the research had achieved a grade A or B in the live[1] examination. The decisions from the silent and verbalising conditions were statistically compared.

Findings

Our interim findings suggest that verbalising made little difference to the participants' decisions; this is in line with previous research in other contexts. The findings reassure us that the verbal protocols are a useful method for research about decision making in both marking and grading.

Background

The 'think aloud' method entails people verbalising their thoughts while they perform tasks. The resulting 'verbal protocols' are then analysed by researchers. The think aloud procedure is an established method of researching what people pay attention to, or what cognitive strategies they are using when they do various complex tasks (e.g. Van Someren

[1] Live is used to denote the examination or procedures taking place 'for real' rather than as part of an experimental setting.

et al., 1994; Taylor and Dionne, 2000). Verbal protocols have been widely used to investigate decision making processes in educational assessment (Cumming, 1990; Sanderson, 2001). Various studies carried out by Cambridge Assessment used verbal protocols to investigate the judgement process involved in marking varied A-level and GCSE examinations (Suto and Greatorex, 2008a and b; Crisp, 2007 and 2008a), as well as to explore the process of judging grading standards in A-levels (Crisp, 2007 and 2008b). One frequent question to Crisp, Suto and Greatorex from researchers and assessment professionals was whether the method of verbalising alters the outcomes of decision processes. Researchers have studied to what extent verbalising affected the cognitive processes in various settings (e.g. Ummelen and Neutelings, 2000). Although the research is currently inconclusive most results show longer performance times and no alternative task outcome (Krahmer and Ummelen, 2004).

There is one piece of research which answers our question in the context of A-level *marking*. Crisp (2008c) collected decision outcomes from two conditions:

1. marking silently;
2. marking whilst thinking aloud.

The mark to re-mark differences in the two conditions were similar. However, it is important to confirm whether verbalising affects decisions about *grading* standards because marking and grading are two distinct but linked procedures in the context of A-level and GCSE examinations. Grading (awarding) meetings involve senior examiners recommending grade boundaries after marking has been completed.

In this article we aim to answer the frequently asked question of whether thinking aloud results in different decisions about A-level grading standards. Ensuring the robustness of research on the psychology of decision making processes in assessment is of crucial importance, especially when using the think aloud method. Therefore, our article draws from a wider project where the main aim was to find out more about cognitive decision making processes used to make judgements about grading standards. As well as studying the decision making processes, the outcomes of the decisions were also considered to be important.

In the wider project five senior examiners[2] made decisions about A-level grading standards for a science examination both silently and whilst thinking aloud. All the decisions were made in experimental conditions for research purposes. Three approaches are considered:

i. awarding – part of the conventional approach to recommending grade boundaries,
ii. Thurstone pairs,
iii. rank ordering.

The latter two were suggested as possible future methods of recommending grade boundaries by Pollitt and Elliott (2003a and b), and Black and Bramley (2008). They have also been used in a series of comparability studies (e.g. Forster and Gray, 2000; Arlett, 2003; Greatorex *et al.*, 2002, 2003; Edwards and Adams, 2002, 2003; Guthrie, 2003; Bramley *et al.*, 1998; Townley, 2007; Black and Bramley, 2008). The focus of this article will be a statistical comparison of the decisions from the silent and verbalising conditions, and is the first in a series of linked studies which make up the wider project. So far only one other study from the wider project is complete, in which Greatorex *et al.* (2008) analysed which items the participants attended to whilst making decisions and whether these items were likely to be helpful in decision making.

In this research we focus on one decision-making phase of awarding which involves the awarding committee judging whether a small number of examples of candidates' work on particular marks show the distinguishing characteristics of performance at a particular grade. For a fuller description see Cresswell (1997), QCA (2008) or Greatorex (2003). The candidates' work is usually examination scripts but might be a recording of a drama or musical performance or an artefact such as a painting. Thurstone pairs and rank ordering, as well as examples of their use in comparability studies, have been frequently described in the literature; see for example Bramley *et al.* (1998), Arlett (2003), Greatorex *et al.* (2002, 2003), Edwards and Adams (2002, 2003), Guthrie (2003) and Townley (2007). Therefore, we will only provide a summary here. Thurstone pairs and rank ordering involve a group of experts judging the quality of candidates' work. In Thurstone pairs in this context each expert compares a pair of scripts, with each pair constituting a script from the live examination and the archive examination. Each expert decides which of two scripts shows evidence of better candidate performance, without re-marking the scripts. This is repeated for a variety of pairs of scripts. When all the necessary comparisons have been made, they are statistically analysed (using Rasch). The results of the analysis can be used to identify a small range of marks within which the live boundary should lie for the standard from last year to be maintained. In rank ordering each expert receives small samples of live and archive scripts which they rank according to the candidates' performance. This is repeated for a number of overlapping samples of scripts. The outcomes of the rankings are submitted to the same statistical analysis as above. Again the statistics can be used to identify a small range of marks within which the live boundary should lie.

There are a number of aspects of awarding meetings and scripts that positively and negatively influence judgements of gradeworthiness (Cresswell, 1997; Murphy *et al.*, 1995; Crisp, 2007; Baird, 2000; Baird and Scharaschkin, 2002; Scharaschkin and Baird, 2000). To understand some of the resulting difficulties we have to bear in mind that A-level and GCSE examinations have a principle of compensation, according to which candidates gain marks for their strengths, and there is more than one way to achieve a grade. Arguably, one issue influencing examiners' grading decisions is that sometimes the visibility of marks given to candidates' responses and the marks available on the question paper become extraneous information. Two conundrums relate to the principle of compensation and the visibility of marks on scripts:

- Some awarding committee members pay particular attention to questions and marks which are believed to differentiate between performances at particular grades (Murphy *et al.*, 1995; Greatorex *et al.*, 2008). This belief might be well or ill founded (Murphy *et al.*, 1995). Focussing on particular questions at the expense of other questions is not aligned with the principle of compensation[3].

[2] All participants in the wider project had been involved in the live Award for the examination and met the criteria used in many comparability studies for recruiting participants.

[3] Grade descriptors are a written summary of the features of performance at particular grades. It is important to note these are indicators of typical performance and not criteria to be met. The grade descriptors are cues to memory which can be used in Awarding meetings. Some might argue that GCSEs or A-levels did not have a principle of compensation because some grade descriptors, including those for the science examination in the research, refer to some high grade performances as being consistently high achieving. However, the principle of compensation holds as these are indicators not criteria, so students do not have to have all the characteristics in the grade descriptors to get the grade. For more details about grade descriptors see Greatorex (2001, 2002, 2003b).

Psychological research from a variety of contexts presented by Greatorex (2007) and later Greatorex *et al*. (2008) suggests that humans are not particularly good at combining information to make decisions. Therefore, focussing judgements on particular questions might be a successful approach to decision making, if the questions are a good proxy for the whole of the examination. After all, the alternative strategy – judgements about whole scripts – involves mentally combining an examinee's answers to all questions in the examination.

- It has been established that the consistency of candidates' performance across questions on an examination paper influences the severity of judgements of gradeworthiness (Cresswell, 1997; Scharaschkin and Baird, 2000). Again, this is not aligned with the principle of compensation.

Given that, arguably, marks can act as extraneous information and that scripts are cleaned of marks in some comparability studies, we decided to include the visibility of marks as a variable in our research.

How reliable are the judgements made using each method?

This research also provided an opportunity to compare the reliability of judgements made under a variety of conditions by using the different methods mentioned above and by adding the visibility of marks as a variable. The reliability of awarding judgements is a well researched topic. As has been apparent for some time, the precision of awarding is less than perfect. For example, see Willmott and Nuttall's (1975) work about the General Certificate of Education O Levels and the Certificate of Secondary Education, the predecessor qualifications of GCSE. In later research, Good and Cresswell (1988) replicated some awarding meetings for French, History and Physics. Good and Cresswell (1988; p. 23 in Cresswell, 2000) concluded that 'different groups of grade Awarders can reach decisions about final grade boundaries which are sufficiently similar to be acceptable, given the inherent imprecision of the examining process'. They aimed to find out what percentage of candidates' grades would have changed if one awarding team's judgements were substituted for another's. They found that 13% of candidates' grades would have changed in French, 17% in Physics and 38% in History. This finding might raise questions about the reliability of awarding procedure; however, Cresswell (2000) cites Willmut (1981) who showed that the change of 38% of grades outcomes corresponds approximately to an inter-examiner reliability coefficient of 0.96, which is generally considered to be of very high inter-rater reliability. Previous research on awarding has established that the severity of judgements of gradeworthiness can be influenced by several factors of arguably varying validity. For instance:

i. the archive scripts provided (Baird, 2000),

ii. the consistency of performance in scripts (Scharaschkin and Baird, 2000) and

iii. whether the examiners see candidates' work from one examination or the work of candidates from the whole qualification (Baird and Scharaschkin, 2002).

In summary, experiments suggest that awarding committees do not make perfectly consistent decisions.

There is little research in the public domain comparing the reliability of decisions made in studies using Thurstone pairs and rank ordering with other approaches to assessment. The main body of evidence in the public domain is by Kimbell *et al*. (2007), who used a mixture of Thurstone pairs and rank ordering to assess a series of Design and Technology portfolios in a pilot study. They claim that their approach to assessment is highly reliable (Kimbell *et al*., 2007). Of course, there is already a large literature illustrating that the reliability of marking is generally less than perfect, for example, see Hartog and Rhodes (1935), Pillner (1968), Willmott and Nuttall (1975), Newton (1996), Pinot de Moira *et al*. (2002), Raikes and Massey (2007) and Vidal Rodeiro (2007). Overall the research is inconclusive regarding whether Thurstone pairs/rank ordering decisions are of similar reliability to more conventional approaches to making assessment decisions. The present research adds to the accumulation of evidence.

Method

Verbal protocols result from participants verbalising their thoughts as, or after, they perform a complex cognitive activity. This is an established method of studying what people pay attention to, or what strategies they use when they are undertaking a variety of complex cognitive tasks (e.g. Van Someren *et al*., 1994; Taylor and Dionne, 2000), including decisions in educational assessment (Crisp, 2008a and b; Suto and Greatorex, 2008a and b; Green 1998; Cumming, 1990; Vaughan, 1992; Weigle, 1994; Milanovic *et al*., 1996).

One of the most established approaches to using verbal protocols is explained by Ericsson and Simon (1993). The approach is also recommended by Krahmer and Ummelen (2004) because it has a sound theoretical basis underpinning it, which some rival approaches do not. The thinking aloud procedure in this research reflected Ericsson and Simon's principles. For instance, the participants had a practice session, and were not interrupted whilst providing the 'real' verbal protocol. The exception to this principle was when a participant was silent for some time and the researcher said 'please keep talking'. The participants were asked to say which script and item they were looking at in the verbal protocols to facilitate the analysis.

Initially, the participants made awarding, Thurstone pairs and rank ordering judgements silently at home. The tasks were then repeated whilst thinking aloud as the main data collection phase. There were some differences between the script samples and procedures for the decisions made silently and whilst thinking aloud (more details are given later). This was because there were only a limited number of scripts to work with and the arrangements for the main data collection phase took precedence over the arrangements for the decisions made silently.

Examination

An AS-level science examination from 2005 and another from 2006 were used in the research. The examinations were from the same qualification and specification. The candidates' work is likely to provide evidence of numerical skills, written skills, use of diagrams and knowledge and understanding. Therefore, research results from this examination might be more generalisable than those from a different subject.

For each examination a total of 45 marks were available. In the live examination the question papers were given to candidates as a form in which the items and source material (e.g. diagrams) were presented along with an answer space into which they added their responses. All the items were worth between 1 and 6 marks. Additionally, one mark was available on each question paper for QWC (quality of written communication) and this was associated with one item on each paper worth 6 marks (plus 1 mark for QWC). The mark scheme was a points based mark scheme.

Script samples

The script samples constituted scripts with total marks within the range of marks considered in the recommendation for the grade A boundary in the awarding meeting (33 to 37 for 2005 and 28 to 34 for 2006). The live grade A boundary was 35 marks for the 2005 examination and 31 marks for the 2006 examination. The frequency of scripts in the sample for the decisions made whilst thinking aloud and the decisions made silently are given in Table 1.

Table 1: Frequency of the scripts with a particular mark

Total marks from 2005	*frequency of scripts in the silent conditions*	*frequency of scripts in the thinking aloud conditions*	***Total marks from 2006***	*frequency of scripts in the silent conditions*	*frequency of scripts in the thinking aloud conditions*
33	2	3	**28**	0	3
34	2	3	**29**	0	5
35	2	7	**30**	2	5
36	2	3	**31**	2	3
37	2	3	**32**	2	4
			33	2	5
			34	2	4
Total	**10**	**19**		**10**	**29**

Participants

Five senior examiners who were involved in recommending live grade boundaries for the AS-level examination in either 2005 and/or 2006 took part in the research.

Conditions

The awarding conditions reflected the aspect of awarding where individual awarding committee members evaluate scripts, prior to coming to a collective view about where the grade boundary should be. The rank ordering and Thurstone pairs conditions were intended to reflect current/best practices in previous studies. For all conditions some minor adjustments were made to current/best practices for the purposes of this research (e.g. asking participants to think aloud).

In our study photocopies of the scripts were used rather than the original scripts. For each method the scripts were presented as they are normally presented: awarding with marks visible, Thurstone pairs with marks visible[4] and rank ordering with scripts cleaned of marks. For awarding and Thurstone pairs the procedures were also undertaken with the scripts cleaned of marks. This experimental control was introduced given the arguably extraneous influence of visible marks in some awarding judgements (Murphy *et al*., 1995; Cresswell, 1997; Scharaschkin and Baird, 2000).

This gave us 10 different experimental conditions:

Table 2: Experimental conditions

Awarding method	*Scripts cleaned of marks (clean) or with marks visible (visible)*	*Decisions made silently (silent) or whilst thinking aloud (VP)*	*Term to be used to refer to the condition*
awarding	Clean	silent	awarding clean silent
awarding	Clean	VP	awarding clean VP
awarding	Visible	silent	awarding visible silent
awarding	Visible	VP	awarding visible VP
rank ordering	Clean	silent	rank ordering clean silent
rank ordering	Clean	VP	rank ordering clean VP
Thurstone pairs	Clean	silent	Thurstone pairs clean silent
Thurstone pairs	Clean	VP	Thurstone pairs clean VP
Thurstone pairs	Visible	silent	Thurstone pairs visible silent
Thurstone pairs	Visible	VP	Thurstone pairs visible VP

Guarding against order effects

Scripts were included in more than one condition when the decisions were made silently. However, each participant undertook the tasks in a different order to guard against order effects[5].

For the main data collection phase each participant experienced the conditions one after the other in the Cambridge offices with a researcher present. (Unfortunately, sometimes the participants did not complete all the tasks due to time constraints and therefore there were some missing data). Three precautions were followed to minimise order effects:

- each participant experienced each condition in a particular order;
- in between undertaking one verbal protocol condition and the next the participants took a break or undertook a distractor task. In the distractor task the participants considered some examination questions from an international syllabus (in the same school subject) at a lower level and rated how accurately they thought different groups of examiners would mark the questions;
- for any given participant each script only appeared in one condition; this was also to guard against participants remembering the scripts.

Additionally, the scripts were designated across tasks and participants to avoid interactions.

For all the conditions, the instructions used for making decisions silently were similar to those used in the main data collection phase. For all conditions, only the question paper, scripts and mark scheme were available for reference. The participants were asked not to use the mark scheme to re-mark the scripts. (The additional information that is provided in live awarding meetings was not provided in this research as it might have influenced the judgements in the other conditions).

For the main data collection phase the instructions used in the Thurstone pairs and rank ordering conditions closely resembled the instructions from the most recent Cambridge Assessment studies in the

[4] Scripts with marks visible have been used in most of the recent inter-Awarding Body comparability studies for UK examinations. These studies were conducted using Thurstone pairs. To explain why cleaning scripts of marks was not necessary in these studies we need to consider what the participants were doing. The participants were asked to make comparisons at the qualification rather than the examination level. That is they would be comparing say three scripts from one candidate who took AQA A-level Chemistry and another three scripts from another candidate who took OCR A-level Chemistry. For a participant to work out a candidate's overall qualification mark they would need to take into account the proportion of the available marks achieved, uniform mark scale calculations as appropriate, the weighting applied to each examination to provide the final overall qualification grade and so on. Given this complexity it is arguably harder to work out how to compare the performances based solely on which candidate has achieved the higher proportion of marks than to make a qualitative judgement about the quality of the performance.

[5] Order effects in this research could be the order in which the conditions and/or scripts were experienced and thereby affecting the decisions.

public domain (Pollitt and Crisp, 2004; Black and Bramley, 2008) with any necessary changes in details for the purposes of this study (e.g. thinking aloud). This was to ensure that current/best practices were followed and to ensure the instructions matched those generally used in studies as deviations from usual practices which would invalidate the research.

The verbal protocols were digitally recorded with the permission of the participants. Subsequently, the digitally recorded information was transcribed.

Analysis

For each awarding condition, the proportion of occasions on which 2006 scripts with a particular mark were judged worthy of a higher grade was calculated. For Thurstone pairs and rank ordering, we calculated the proportion of occasions on which 2005 scripts on a particular mark were judged to be better than scripts from 2006 on a particular mark. For instance, we calculated the proportion of occasions on which 2005 scripts with a mark of 35 were judged to be better than (winning against[6]) 2006 scripts of 29 marks. The calculations were undertaken for each mark from each year. The figures were calculated separately for the scripts with marks visible and the scripts cleaned of marks. (Note that for rank ordering the experts ranked two samples of scripts and the outcomes could not be statistically combined so the results have been presented separately.)

The resulting patterns from the decisions made silently and whilst thinking aloud were compared by scanning the figures. The figures are indicated in Tables 3 to 5. We predicted we would gain an approximate increasing monotonic relationship from the bottom left corner to the top right corner in each table because we expected that:

- In the awarding conditions 2006 scripts with higher marks will be judged worthy of a higher grade on a higher proportion of occasions. For example, 28-mark scripts (from 2006) should be judged worthy of grade A on a smaller proportion of occasions than 37-mark scripts (from 2006).
- In the Thurstone pairs and rank ordering conditions the proportion of occasions on which 2005 scripts are judged better than 2006 scripts will increase as the 2005 total mark increases. For example, in any comparison 33-mark scripts (from 2005) should win against 28-mark scripts (from 2006) on a smaller proportion of occasions than 37-mark scripts (from 2005) should win against 28-mark scripts (from 2006). Also, in any comparison 33-mark scripts (from 2005) should win against 34-mark scripts (from 2006) on a smaller

[6] When a script is judged to be better in quality than another we sometimes refer to this as 'winning against' another script.

Table 3: The proportion of occasions on which 2006 scripts with a particular mark were judged worthy of grade A in the awarding conditions

Silent

Proportion of occasions on which 2006 scripts were judged worthy of grade A

2006 Mark	*30*	*31*	*32*	*33*	*34*
1					**v**
0.9					
0.8				**v**	
0.75			**v**		
0.6	**c**	**c**			
0.5					
0.4				**c**	**c**
0.3					
0.25		**v**	**c**		
0.1					
0	**v**				

Thinking Aloud

Proportion of occasions on which 2006 scripts were judged worthy of grade A

2006 Mark	*28*	*29*	*30*	*31*	*32*	*33*	*34*
1					**c**		**v**
0.9							
0.8						**v**	
0.75				**c**			**c**
0.6					**v**		
0.5							
0.4							
0.3				**v**			
0.25		**c**	**c**			**c**	
0.1							
0	**v, c**	**v**	**v**				

c = Cleaned **v** = Visible All marks in the grid are from 2006

How to use Table 3
As an example of how to use Table 3 we can see that the proportion of occasions that 32-mark scripts from 2006 (found in the top row) were judged worthy of grade A for decisions made silently was 0.25 (found in the left hand column) for the cleaned of marks condition (indicated by a **c** in the grid) and 0.75 for the marks visible condition (indicated by a **v** in the grid).

In Table 3 we would expect to get an approximate increasing monotonic relationship from the bottom left hand corner to the top right hand corner. This is because the higher the 2006 mark the greater the proportion of occasions on which scripts should be judged worthy of grade A, irrespective of the visibility of their marks.

proportion of occasions than 37-mark scripts (from 2005) should win against 34-mark scripts (from 2006).

All the marks shown in the analysis refer to the total marks achieved by candidates on the examination paper in question.

According to some awarding committee members it is difficult to make decisions about 'rogue' or apparently atypical scripts, and they usually avoid using such scripts in making recommendations for grade boundaries in live awarding meetings, for example, Murphy *et al*. (1995). However, it was assumed for this study that all scripts in the sample on a particular mark had the characteristics of performance at that mark. This is a reasonable assumption given that in live contexts all scripts on a particular mark are awarded a particular grade.

Results

See Tables 3, 4 and 5.

Findings given in Table 3

Broadly speaking there is little difference between the pattern of decisions made silently in comparison with the pattern of decisions made whilst thinking aloud. This reinforces the findings in previous literature.

However, there was a considerable difference between the pattern of decisions made with scripts with marks visible and the pattern of decisions made on scripts cleaned of marks. When the marks were visible the expected pattern was evident. We can see that the expected pattern was not anywhere near as clear for decisions made when the scripts were cleaned of marks.

Findings given in Table 4

Overall, in general, the expected pattern is evident, for decisions made:

i. silently and whilst thinking aloud,

ii. with scripts with marks visible and scripts cleaned of marks.

However, there are some scripts which do not conform to the pattern.

Findings given in Table 5

For both decisions made silently and whilst thinking aloud the broad pattern is similar; in both cases as the 2005 marks increase the proportion of occasions on which the 2005 scripts win also increases. However, there are a few scripts which seemed to be ranked lower or higher than would be expected given the mark achieved by the candidate.

Table 4: The proportion of occasions on which 2005 scripts won against 2006 scripts in the Thurstone pairs conditions

Silent

Proportion of occasions on which 2005 scripts won against 2006 scripts

2005 Mark	*33*	*34*	*35*	*36*	*37*
1				**30v**	**34c, 31v**
0.8		**30v, 32v**	**32v**	**32c, 31v**	**34v**
		30c	**31c**		
0.6	**34v**		**33v**	**34c**	**33c**
0.4	**32c, 33c**				
	33v				
0.2					
0					

Thinking Aloud

Proportion of occasions on which 2006 scripts were judged worthy of grade A

2005 Mark	*33*	*34*	*35*	*36*	*37*
1	**29v, 32v**	**28c, 33v, 34v**	**29c, 32c**	**30c, 29v, 30v**	**28c, 31c**
		30v	**29v**		**31v, 32v, 33v**
0.8					
0.6					
	30c, 30v	**34c**	**34v**		**33c**
0.4					
0.2					
	31v				
0	**29c, 32c, 34c**	**29c, 31c**	**33v**	**29c, 33c**	**32c**

c = Cleaned **v** = Visible

How to use Table 4

As an example of how to use Table 4 in the marks visible condition the proportion of occasions on which 2005 scripts with 34 marks (found in the top row) won against 2006 scripts with 30 marks (found in the grid) was 0.8 (found in the left hand column), also the proportion of occasions on which 2005 scripts with 34 marks won against 2006 scripts with 32 marks was 0.8.

In Table 4 we expect to see an approximate increasing monotonic relationship from the bottom left hand corner to the top right hand corner. This is because the proportion of occasions on which 2005 scripts should win against 2006 scripts should increase as the 2005 marks increase. The higher the 2005 marks, the larger the proportion of occasions on which 2005 scripts should win against 2006 scripts, irrespective of the 2006 script mark. For example, in any comparison 33-mark scripts (from 2005) should win on a lower proportion of occasions against 28-mark scripts (from 2006) than 37-mark scripts (from 2005) should win against 28-mark scripts (from 2006). Also, in any comparison 33-mark scripts (from 2005) should win on a lower proportion of occasions against 34-mark scripts (from 2006) than 37-mark scripts (from 2005) should win against 34-mark scripts (from 2006).

Table 5: The proportion of occasions on which 2005 scripts won against 2006 scripts in the rank ordering conditions

Silent

Proportion of occasions on which 2005 scripts won against 2006 scripts

2005 Mark	*33*	*34*	*35*	*36*	*37*
1	30				29
			28*		
0.9					
0.8		29, 33	29	29, 30	30
0.7					
	29, 28*	26*		28*	28*
0.6	33	30, 32	30, 33		32, 33
0.5					
0.4	32		32	33	
0.3					
0.2				32	
0.1					
0					

Thinking Aloud

Proportion of occasions on which 2006 scripts were judged worthy of grade A

2005 Mark	*33*		*34*		*35*		*36*		*37*	
	Pk 1	*Pk 2*	*Pk 1*	*Pk 2*	*Pk 1*	*Pk 2*	*Pk 1*	*Pk 2*	*Pk 1*	*Pk 2*
1		30	33	30	31				30	30
				33	33				31	33
									33	
									34	
0.8				32	30	30		30		31
				34						32
										34
0.6			30		34	32		32		
			34			33		34		
						34				
0.4	30	31	31	31	32	31	30	31	32	
	31	32					31			
	31	33					32			
		34					32			
0.2	32		32				34			
0										

* less than 5 participants

How to use Table 5

As an example of how to use Table 5 we can see that the proportion of occasions when 2005 scripts of 33 marks (in the top row) from pack 2 (in the second row from the top) won against 2006 scripts of 30 marks (in the grid) was 1 (in column second from the left); this means the 33-mark 2005 scripts always won.

In Table 5 we would expect to get an approximate increasing monotonic relationship from the bottom left hand corner to the top right hand corner. This is because the proportion of occasions on which 2005 scripts should win against 2006 scripts should increase as the 2005 mark increases, irrespective of the 2006 script mark. For example, in any comparison 33-mark scripts (from 2005) should win against 28-mark scripts (from 2006) on a smaller proportion of occasions than 37-mark scripts (from 2005) should win against 28-mark scripts (from 2006). Also, in any comparison 33-mark scripts (from 2005) should win against 34-mark scripts (from 2006) on a smaller proportion of occasions than 37-mark scripts (from 2005) win against 34-mark scripts (from 2006).

Discussion

Limitations

The experimental conditions reflected current/best practices for awarding, rank ordering and Thurstone pairs procedures within the restrictions of a research study. The first limitation was that the awarding conditions slightly digressed from the awarding practice in two ways:

i. Participants in the experiment did not have any information in addition to scripts that would usually be available in the awarding meeting (apart from the archive scripts, question paper and the mark scheme). This was to avoid influencing the decisions made in the other conditions, which do not include using such information.

ii. Awarders do not always individually make decisions about the quality of candidates' work, although this is not uncommon (Cresswell, 1997). Individual rather than collaborative decisions about individual scripts might increase *if* awarding meetings were undertaken remotely.

Therefore, the awarding conditions in this study might have somewhat limited ecological validity for decisions made silently as well as those made whilst thinking aloud.

A second limitation of our analysis was that the design of the study focused on the main aim of a wider project – to know more about how decisions about grading standards are made from a psychological perspective – and the purpose of the current analysis was of less importance in the study design. There are several aspects to this limitation:

i. The robustness of the statistics is compromised by the small samples of participants and scripts which also affects the generalisability of the study.

ii. The scripts judged silently and whilst thinking aloud were mutually exclusive samples of scripts, consequently, any differences between the modes of grading could be due to the different script samples. The results indicated that the decisions made silently and whilst thinking aloud were broadly similar and therefore this design limitation did not seem to affect the results.

Overall findings

Broadly speaking, verbalising made little difference to the participants' decisions in the various experimental conditions. This is in line with previous research about decisions made silently and whilst thinking aloud in a variety of contexts. Crisp's finding (2008c) that the *marking* decisions made silently are broadly similar to the decisions made whilst thinking aloud is of particular importance to the assessment community. Our findings are in line with those of Crisp (2008c); these studies reassure us that think aloud concurrent verbal protocols are a robust method for research on decision making in both *marking* and *grading*.

Thus far there is little research literature in the public domain regarding whether the decisions made in Thurstone pairs exercises, rank ordering studies or awarding are the most reliable. The literature also indicates that the visibility of marks can affect decisions in awarding meetings, although there is no similar research for Thurstone pairs and rank ordering. In the present study for all conditions we expected to see an approximate increasing monotonic relationship. For awarding conditions we expect the proportion of occasions on which 2006 scripts are judged worthy of grade A to increase as the 2006 marks increase. For Thurstone pairs and rank ordering conditions we expect the proportion of occasions on which 2005 scripts win against 2006 scripts to increase as the 2005 marks increase. The statistics from the present study suggest that:

- When the scripts are cleaned of marks, the participants make decisions along the expected pattern in Thurstone pairs and rank ordering, but this was not true for the awarding conditions.
- Decisions follow the expected pattern in the Thurstone pairs and awarding conditions with the marks visible. The later can be used to argue that awarding judgements are highly reliable, despite the research literature indicating that the reliability of awarding judgements is less than perfect. On the other hand, it can be argued that given the research literature and the findings of this research, the pattern is almost 'too perfect' for the awarding conditions with marks visible. Perhaps this indicates that the participants relied heavily on the visibility of marks to make their decisions rather than the contents of the scripts or the quality of the candidates' performance.

Laming (2004) theorises that generally, people can make comparisons between two artefacts, but they are not able to maintain a standard in mind and use it to make consistent decisions. This explains why the participants were arguably better at making Thurstone pairs and rank ordering judgements (comparisons between scripts) than making awarding judgements (comparing scripts with internal standards) when the scripts are cleaned of marks. After all, when the marks are visible, decisions can be made based on the marks rather than the examiners' judgements about the gradeworthiness. Indeed, Laming's theory has been used to argue that Thurstone pairs and rank ordering are better methods for maintaining and/or comparing standards than the methods requiring participants to maintain internal standards, for example, at awarding meetings (Bramley, 2005, 2007; Greatorex, 2007; Black and Bramley, 2008).

Implications and recommendations

The present research and other studies (Crisp, 2008c), reassure us that think aloud verbal protocols are a robust research method in the sense that the outcome decisions are unaffected by verbalisation. Therefore, we recommend using concurrent think aloud verbal protocols in future research studies regarding assessment decisions.

This also signifies that our extensive research about the marking and grading judgement processes utilising the method of concurrent think aloud procedure is a trustworthy source of evidence. (See Suto *et al*., 2008) for an overview of the research on the judgement processes in examination marking). The rich qualitative verbal protocol data collected in the main data collection phase of the wider research project are still being analysed and it is hoped that analyses will add to our knowledge about how decisions are made about grading standards.

References

Arlett, S. J. (2003). *A Comparability Study in VCE Health and Social Care, Units 3, 4 and 6: A review of the examination requirements and a report on the cross moderation exercise. A study based on the Summer 2002 examination and organised by AQA on behalf of the Joint Council for General Qualifications*.

Baird, J. (2000). Are examination standards all in the head? Experiments with examiners' judgements of standards in A level examinations. *Research in Education*, **64**, 91–100.

Baird, J. & Scharaschkin, A. (2002). Is the Whole Worth More than the Sum of the Parts? *Studies of Examiners' Grading of Individual Papers and Candidates' Whole A-Level Examination Performances*. *Educational Studies*, **28**, 2, 143–162.

Black, B. & Bramley, T. (2008). Investigating a judgemental rank ordering method for maintaining standards in UK examinations. *Research Papers in Education*, **23**, 3, 357–373.

Bramley, T. (2005). A Rank-Ordering Method for Equating Tests by Expert Judgement. *Journal of Applied Measurement*, **6**, 2, 202–223.

Bramley, T. (2007). Paired Comparison Methods. In: P. Newton, J. Baird, H. Goldstein, H. Patrick & P. Tymms (Eds.), *Techniques for monitoring the comparability of examination standards*. (pp. 246–294) QCA: London.

Bramley, T., Bell, J. F. & Pollitt, A. (1998). Assessing changes in standards over time using Thurstone paired comparisons. *Education Research and Perspectives*, **25**, 2, 1–23.

Cresswell, M. (1997). *Examining Judgements: Theory and Practice of awarding public examination grades*. PhD thesis, University of London Institute of Education: London.

Cresswell, M. (2000). Defining, Setting and Maintaining Standards in Curriculum-Embedded Examinations: Judgemental and Statistical Approaches. In: H. Goldstein & T. Lewis (Eds.), *Assessment: Problems, developments and statistical issues*. (pp. 57 to 84). Chichester: John Wiley and Sons.

Crisp, V. (2007). *Do assessors pay attention to appropriate features of student work when making assessment judgements?* A paper presented at the International Association for Educational Assessment Annual Conference, Baku, Azerbaijan, September, 2007.

Crisp, V. (2008a). Exploring the nature of examiner thinking during the process of examination marking. *Cambridge Journal of Education*, **38**, 2, 247–264.

Crisp, V. (2008b). *Judging the grade: An exploration of the judgement processes involved in A level grading decisions*. A paper presented at the British Educational Research Association Conference, September, Heriot-Watt University.

Crisp, V. (2008c). The validity of using verbal protocol analysis to investigate processes involved in examination marking. *Research in Education*, **79**, 1, 1–12.

Cumming, A. (1990). Expertise in evaluating second language compositions. *Language Testing*, **7**, 31–51.

Edwards, E. & Adams, R. (2002). *A Comparability Study in GCE Advanced Level Geography Including the Scottish Advanced Higher Grade Examinations: A review of the examination requirements and a report on the cross moderation exercise. A study based on the Summer 2001 Examination and organised by WJEC on behalf of the Joint Council for General Qualifications*.

Edwards, E. & Adams, R. (2003). *A Comparability Study in GCE Advanced Level Geography Including the Scottish Advanced Higher Grade Examinations: A review of the examination requirements and a report on the cross moderation exercise. A study based on the Summer 2002 Examination and organised by WJEC on behalf of the Joint Council for General Qualifications*.

Ericsson, K. & Simon, H. (1993). *Protocol Analysis: Verbal reports as data*. Cambridge, MA: MIT Press.

Forster, M. & Gray, E. (2000). *Impact of Independent Judges in comparability studies conducted by Awarding Bodies*. A paper presented at the British Educational Research Association Annual Conference, Cardiff University, September.

Good, F. J. & Cresswell M. J. (1988). Grading the GCSE. London: Secondary schools Examination Council. In: M. Cresswell (2000) *Defining, Setting and Maintaining Standards in Curriculum-Embedded Examinations: Judgemental and Statistical Approaches*. In: H. Goldstein & T. Lewis, (Eds.) *Assessment: Problems, developments and statistical issues*. (pp. 57–84). Chichester: John Wiley and Sons.

Greatorex, J. (2003). *What happened to limen referencing? An exploration of how the Awarding of public examinations has been and might be conceptualised*. A paper presented at the British Educational Research Association Conference, 10–13 September 2003 at Heriot-Watt University, Edinburgh.

Greatorex, J. (2007). *Contemporary GCSE and A-level Awarding: A psychological perspective on the decision-making process used to judge the quality of candidates' work*. A paper presented at BERA 2007, University of London.

Greatorex, J., Elliott, G. & Bell, J. F. (2002). *A Comparability Study in GCE AS Chemistry Including parts of the Scottish Higher Grade Examinations: A review of the examination requirements and a report on the cross moderation exercise. A study based on the Summer 2001 examination and organised by the Research and Evaluation Division, University of Cambridge Local Examinations Syndicate for OCR on behalf of the Joint Council for General Qualifications*.

Greatorex, J., Hamnett, L. & Bell, J. F. (2003). *A Comparability Study in GCE Chemistry Including the Scottish Advanced Higher Grade. A study based on the Summer 2002 examination and organised by the Research and Evaluation Division, University of Cambridge Local Examinations Syndicate for OCR on behalf of the Joint Council for General Qualifications*.

Greatorex, J., Novakovic, N. & Suto, I. (2008). *What attracts judges' attention? A comparison of three grading methods*. A paper presented at the IAEA conference, Cambridge.

Green, A. (1998). *Studies in language testing 5: verbal protocol analysis in language testing research*. Cambridge: Cambridge University Press.

Guthrie, K. (2003). *A Comparability Study in GCE Business Studies and VCE Business: A review of the examination requirements and a report on the cross moderation exercise. A study based on the Summer 2002 Examination and organised by the EdExcel on behalf of the Joint Council for General Qualifications*.

Hartog, P. & Rhodes, E. C. (1935). *An Examination of Examinations*. London: Macmillan.

Kimbell, R., Wheeler, A., Miller, S. & Pollitt, A. (2007). *E-scape portfolio assessment phase 2 report*. Department of Design, Goldsmiths, University of London. http://www.goldsmiths.ac.uk/teru/UserFiles/File/e-scape2.pdf

Krahmer, E. & Ummelen, N. (2004). Thinking about thinking aloud: A comparison of two verbal protocols for usability testing. *IEEE Transactions of Professional Communication*, **47**, 2, 105–117.

Laming, D. (2004). *Human judgment: The Eye of the Beholder*. London: Thomson.

Milanovic, M., Saville, N. & Shuhong, S. (1996). A study of decision-making behaviour of composition markers. In: M. Milanovic & N. Saville (Eds.), *Studies in Language Testing 3*. Cambridge: Cambridge University Press.

Murphy, R., Burke, P., Cotton, T., Hancock, J., Partington, J., Robinson, C., Tolley, H., Wilmut, J. & Gower, R. (1995). *The dynamics of GCSE Awarding*. Report of a project conducted for the School Curriculum and Assessment Authority, School of Education, University of Nottingham.

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

Pillner, A. E. G. (1968). Examinations. In: H. J. Butcher (Ed.), *Education Research in Britain*, pp. 167–184. London: University of London Press.

Pinot de Moira, A., Massey, C., Baird, J. & Morrissy, M. (2002). Marking consistency over time, *Research in Education*, **67**, 79–87.

Pollitt, A. & Crisp, V. (2004). *Could comparative judgements of script quality replace traditional marking and improve the validity of examination questions?* A paper presented at the British Educational Research Association, Conference, Manchester.

Pollitt, A. & Elliott, G. (2003a). *Monitoring and Investigating comparability: a proper role for human judgement*. Qualifications and Curriculum Authority Seminar on 'Standards and Comparability', UCLES 4th April 2003.

Pollitt, A. & Elliott, G. (2003b). *Finding a proper role for human judgement in the examination system*. Qualifications and Curriculum Authority Seminar on 'Standards and Comparability', UCLES 4th April 2003.

Qualifications and Curriculum Authority (2008.) *GCSE, GCE, and AEA code of practice 2008*. QCA: London.

Raikes, N. & Massey, A. (2007). Item-level examiner agreement. *Research Matters: A Cambridge Assessment Publication*, **4**, 34–37.

Sanderson, P. J. (2001). *Language and Differentiation in Examining at A level*. PhD thesis, School of Psychology, University of Leeds.

Scharaschkin, A. & Baird, J. (2000). The effects of consistency of performance on A Level examiners' judgements of standards. *British Educational Research Journal*, **26**, 3, 343–357.

Suto, W. M. I., Crisp, V., & Greatorex, J. (2008). Investigating the judgemental marking process. *Research Matters: A Cambridge Assessment Publication*, **5**, 6–8.

Suto, W. M. I. & Greatorex, J. (2008a). What goes through an examiner's mind? Using verbal protocols to gain insights into the GCSE marking process, *British Educational Research Journal*, **34**, 2, 213–233.

Suto, W. M. I. & Greatorex, J. (2008b). A quantitative analysis of cognitive strategy usage in the marking of two GCSE examinations. *Assessment in Education: Principles, Policy and Practice*, **15**, 1, 73–89.

Taylor, K. L. & Dionne, J-P. (2000). Accessing problem-solving strategy knowledge: the complementary use of concurrent verbal protocols and retrospective debriefing. *Journal of Educational Psychology*, **92**, 413–425.

Townley, C. (2007). *Australian Education Systems Officials Committee – Secondary Schools Reporting – A study to examine the feasibility of a common scale for reporting all senior secondary subject results*. Victoria Curriculum and Assessment Authority.

Ummelen N., & Neutelings, R. (2000). Measuring reading behavior in policy documents: A comparison of two instruments. *IEEE Trans. Profess. Commun.*, **43**, 3, 292–302. In: E. Krahmer & N. Ummelen (2004), Thinking about thinking aloud: A comparison of two verbal protocols for usability testing. *IEEE Transactions of Professional Communication*, **47**, 2, 105–117.

Van Someren, M., Barnard, Y. & Sandberg, J. (1994). *The think aloud method: a practical guide to modelling cognitive processes*. London: Academic Press.

Vaughan, C. (1992). Holistic assessment: what goes on in the rater's mind? In: L. Hamp-Lyons (Ed.) *Assessing second language writing in academic contexts*. Norwood, NJ: Ablex.

Vidal Rodeiro, C. L. (2007). Agreement between outcomes from different double marking models. *Research Matters: A Cambridge Assessment Publication*, **4**, 28–34.

Weigle, S. C. (1994). Effects of training on raters of ESL compositions, *Language Testing*, **11**, 197–223.

Willmott, A. S. & Nuttall, D. L. (1975). *The reliability of examinations at 16+*. Schools Council Research Studies. Schools Council Publications. London: MacMillan Education Ltd.

Wilmut, J. (1981). *A Brief Report on two factors which Affect Grade Changes in Mark-Remark and Weighted Exercises*. Associated Examining Board Research report. RAC/184. Guildford: AEB. In: M. Cresswell (2000), Defining, Setting and Maintaining Standards in Curriculum-Embedded Examinations: Judgemental and Statistical Approaches. In: H. Goldstein & T. Lewis (Eds), *Assessment: Problems, developments and statistical issues*. pp. 57–64. Chichester: John Wiley and Sons.