Could Comparative Judgements Of Script Quality Replace Traditional Marking And Improve The Validity Of Exam Questions?

Alastair Pollitt and Victoria Crisp

University of Cambridge Local Examinations Syndicate

A paper to be presented at the British Educational Research Association Annual Conference, UMIST, Manchester, September 2004.

Contact details
Victoria Crisp
UCLES
1 Hills Road, Cambridge
CB1 2EU
Tel. 01223 553805
Fax. 01223 552700
Email: crisp.v@ucles.org.uk

Disclaimer

The opinions expressed in this paper are those of the authors and are not to be taken as the opinions of the University of Cambridge Local Examinations Syndicate (UCLES) or any of its subsidiaries

Acknowledgements

The authors would like to thank Stuart Currie for his assistance in adapting questions, Stuart Currie, Janet Wooley, Andy Daly and Cliff Shortell for carrying out the paired comparison exercise, the teachers and students at the school involved for their cooperation and Jane Fidler for her administrative assistance.

Could comparative judgements of script quality replace traditional marking and improve the validity of exam questions?

Abstract

The traditional marking of examination questions makes reliability crucial. This in turn can sometimes lead to the development of questions that do not ask what examiners really wanted to find out whether students could answer. For example, a question might ask for a list of four reasons for something in order to help ensure reliable marking when what setters really wanted to find out was whether candidates could 'explain why'. Certain topics can prove difficult to test validly because of this in the current examination system. This problem seems to occur most in some areas of less traditional subjects (e.g. in Information Technology) for a number of possible reasons: perhaps because there is a less well-defined historically-developed body of knowledge in such subjects, perhaps due to overlap of content with knowledge from everyday experience, or perhaps because sometimes a topic that is usually practical needs to be tested theoretically when this may not be the most suitable means of assessment.

In studies of the comparability of standards in syllabus specifications across different exam boards, Thurstone paired comparisons are used to generate a rank ordering of candidates' scripts. The method involves examiners taking two scripts from a pile, judging the quality of the responses in them and deciding which script is 'better' and then taking another two and doing the same and so on. It has been suggested (Pollitt and Elliott, 2003) that the use of Thurstone paired comparison judgements could avoid the need for traditional marking as it provides a rank ordering of scripts and grading could be done at the same time. This would reduce restrictions on the way that questions are written since questions would not need to be such that they can be marked reliably hence reducing the risks to question validity mentioned above. Instead of counting the number of correct points students make, the method relies on judgement of the comparative quality of responses in entire scripts (or even each student's entire set of work for a subject). Pilot work on using the Thurstone method to allow improvement of question validity will be described.

Introduction

The main aim of summative assessment is to obtain a measure of the knowledge and abilities (or mental traits) of a group of people in a way that generalises usefully to serve certain purposes. There is no necessary reason why this should mean that assessment always results in scores, though this is almost always assumed to be the case. Even the best known text books on test theories generally assume, without justification, that assessment will result in scores (e.g. Hambleton and Swaminathan, 1985).

What is required of summative assessment is essentially to sort candidates into a rank order to an acceptable degree of accuracy and usually to apply a standard to that order so that the results can be meaningfully understood. Pollitt (2004) defined the fundamental purpose of summative assessment as follows:

> "We are required to judge the overall quality of students (or their performances) in some educational domain on a standard ordinal scale." (Pollitt, 2004, p.4)

This purpose only requires ordinal measurement but it is generally expected that outcomes will be provided on an interval scale to facilitate interpretation by reference to 'standards'.

This paper will focus on one particular problematic issue associated with the predominance of scoring assessments: the need for questions to be such that marking can be reliably carried out by many examiners and the implications of this necessity. Traditional marking of exams demands that the scores awarded are not significantly affected by which marker happens to mark a particular script. In order for public acceptance, marking must be seen to be reliable and fair. This demand is always in the question writer's mind, and has a significant impact on the way that questions are written.

Sometimes examiners cannot set the questions that they want to see students trying to answer – the question has to be distorted so that reliable point-for-point marking can be achieved. The question writing process is constrained because the questions must be able to be marked without much personal judgement on the part of the marker. There is a risk that when setters are prevented from asking the questions that they really want to ask – ones that they think would really address the core knowledge and skills of a subject – less valid questions will result. We have seen some evidence of this from our observations of the question writing and script marking processes and from script analysis. The need for reliable marking sometimes affected the way a question was written, and in turn, the validity of that question.

Take for example the following question that appeared in an Information and Communications Technology exam for sixteen-year-olds.

Describe **three** steps a student will take to find and obtain a picture from the internet.
Step 1 _____
Step 2 _____
Step 3 _____ [3]

The mark scheme credited the following answer.

1. Use a search engine to find a picture.
2. Select or go to one of the sites displayed
3. Copy/ print a picture onto the computer.

Actually, it could be said that there is a whole range of answers to this question that are correct, and students did provide a wide range. For example, it is not entirely clear from the question whether the procedure should start with searching for a picture, or at an earlier stage such as logging on to the internet or deciding what type of image is needed (how large, what file format and quality). It is also not clear whether the student described in the question already has a website in mind where an appropriate image might be found. This might be a feasible interpretation of the question that would lead to the response that the student should start by typing in the address of that website instead of a response referring to using a search engine. In addition, what exactly 'obtain' should be taken to mean is down to interpretation (e.g. a physical print, a copy on file, pasted into a word processing document) although perhaps any interpretation would have been acceptable as point 3 in the marking scheme.

The internet is ultimately a flexible tool and defining the process even of a fairly small task into three set stages seems somewhat simplistic. However, what other choice did the question setter (or question setting team) have to test this area of knowledge when reliable marking was needed? An open question, along the lines of that shown below might have been what they really wanted to find out whether the candidate could respond well to.

Describe how to find and obtain a picture from the internet. [3]

This question still has ambiguities in terms of the exact context of the image search (for example, the purpose that the picture is needed for) but other problems are reduced. Marking such a question reliably would be rather difficult because the points made might have been quite varied and in fact if it was to be marked as one mark for each point made (as is currently usually the case with questions of this length), then it is more honest to point this out by providing labelled response prompts as the examiners did.

Providing a purposeful context for the question (e.g. 'A student needs to find a picture for a school project and insert it into a word processing document') could have improved the question and still provided the three step marking method. This would have reduced

possible variations in responses for the reasons already discussed. However, it would have added substantially to the amount of reading required which may have been unjustifiable in the case of a three-mark question. Additionally, the contextual details might have been distracting and have influenced the kinds of responses elicited (as has been found to sometimes occur, see Ahmed and Pollitt, 2000). The simplest way to ask the question that it seems the question setters wanted to ask would have been the suggestion above. However, this would not suit a point-for-point marking method.

Problems like this seem to occur most prominently in certain less traditional subject areas such as Information and Communications Technology and aspects of Design and Technology and so on. This may be the case for a number of reasons: -

- There is a less well-defined historically-developed body of knowledge in the area so it is difficult to define one right answer to the question that examiners would like to ask, hence leading to amendments.
- Experiences from everyday life will provide knowledge that overlaps with the subject knowledge. This makes the boundaries of the subject knowledge rather fuzzy and leads to some subjectivity over what is good enough to get a mark in an exam.
- The topic is really about practical knowledge but is being tested on paper because of requirements for a certain percentage of assessment to be external. This may not be the most suitable way to assess a particular topic. In the question discussed above, clearly the best way to find out whether a candidate knows how to find and obtain a picture from the internet is to ask them to do it.

It is very difficult to resolve the problems with a question like the one described given the need for reliable marking in an externally assessed test.

In a previous paper, Pollitt (2004) argued that there was an alternative way to assess students without scoring answers and then adding up these marks to get a total score. He argued that the alternative method would closer achieve the fundamental purpose of summative assessment and be intrinsically more valid. Given the defined fundamental purpose, some way of judging performances to create a scale is needed. It seems that concern over the subjectivity of making judgements has tended to lead to the use of the current method of assessment based on scoring many small questions rather than a few larger questions (where this is possible). The current system involves examiners in making microjudgements of performances and these are added up to produce a macrojudgement. However, this method may improve reliability but tends to reduce validity. There is no guarantee that the weighted sum of microjudgements leads to accurate macrojudgements of a student's performance. In the literature on assessment there has been little discussion of judging performances except with regard to essay marking and setting standards (see for example, Wood, 1991). Performances are sometimes judged directly against descriptions of standards in some areas (e.g. art, sport, drama, first and foreign languages, some vocational areas). However, such an approach is unlikely to be generalisable because judgement is always relative – a comparison of two things (Laming, 2004). As a result, making a reliable direct judgement requires remembering or imagining another performance with which to compare and having a series of internalised standards. There are limitations on how many such categories a

person can reliably distinguish: Laming suggests only five. As such, the agreement between raters would be a matter of concern.

Pollitt (2004) argued that since judgements are really comparisons, why not compare performances directly? The alternative measurement method proposed by Pollitt (2004) and previously by Pollitt and Elliott (2003) is based on Louis L Thurstone's Law of Comparative Judgement (Thurstone, 1927), which provides a method of constructing an interval scale from judgements. This is possible because although human judges are likely to have their own internalised standards about what constitutes an item of a certain quality, if they compare two things (as in the Thurstone method) then their own standard cancels out. A true measurement scale can be constructed which shows the value of performances relative to each other. For example, if a particular script is nearly always judged better than the others then it will be fairly high on the scale. The method generates a measurement parameter estimate for each script and also the standard error of that estimate.

The method has been used since 1996 in the field of assessment in comparability studies, looking at the relative standard of different exam boards' assessments in a particular subject. Practically, this has involved a team of examiners reading two scripts (or two sets of student's work), deciding which of the two is 'better', and then comparing another two scripts and so on. Using such a method does not involve any kind of scoring of responses (if the scripts have previously been marked, marks are removed) and hence would not restrict the way that questions are written in order that they can be reliably marked as is currently the case. Questions could be written in a less restricted way and would hence be likely to be more valid. The method relies on judgements of the comparative quality of responses to construct an ordering of candidates instead of on counting the number of correct points made.

In the past the method has not been possible in live examining because each script (or set of work) has to be paired with several other scripts and needs to be judged by more than one examiner. This would not be practical for reasons of time in the current marking system where scripts are delivered by post. However, with electronic script marking (where scripts are scanned and sent electronically to examiners) soon to be a reality, directing one piece of work to several markers would not lead to delay. So the method would involve scripts being sent electronically to judges in pairs and the judges being asked to report which represents a better performance. Each script would be seen several times in different pairings.

Comparability studies using paired comparison judgements have been used in a wide range of subject areas (including history, maths, science, foreign languages, geography and English) which of course use a range of questions types (e.g. short answer, essay, calculations, coursework) and have sometimes involved considering a student's whole set of work for the specification (e.g. six papers). In these studies, the examiners involved have often been initially sceptical about the method, claiming that it seems unnatural to judge scripts containing a number of short answers holistically. However, they all tried the method and most accepted that it could work. The experience so far has demonstrated that the method works in many assessment contexts; however further evidence may well be needed to convince all concerned.

The proposed alternative method has benefits beyond potentially allowing more valid assessment. For example, including some of last year's scripts would allow the standard of the current year's scripts to be determined as part of the judgement process rather than at a separate stage (i.e. award meetings). In addition, scripts that lie close to a boundary between grades and where the standard error goes over the boundary could be sent for extra comparisons to reduce the likelihood of misgrading. The statistical analysis will also pick up misfitting scripts (where there is inconsistency in the judgements about a script). Such scripts, which are proving difficult to judge, could be sent to a senior examiner instead for further judgements. The misfit statistics also allow for the consistency of individual judges to be monitored and could lead to early decisions to stop sending further scripts to an examiner whose judgements are inconsistent with those of the others. These and other potential advantages of the paired comparison method are discussed in greater depth by Pollitt (2004). The main focus of this paper is on the potential benefit to assessment validity.

At award meetings, senior examiners have to make decisions about the total mark that represents the minimum achievement to be awarded a particular grade. At such meetings, a selection of scripts scoring different total marks are considered. It is not unusual when considering several scripts that have achieved identical marks, for examiners to consider some of them more than good enough to achieve the grade, some not quite good enough to achieve the grade and others just good enough to achieve the grade. In order to resolve this, the decision has to be a case of best fit rather than perfect fit. This observation indicates that adding up scores is not a perfect measure of ability or of what students' deserve. The paired comparison method would remove this unfairness and improve validity.

Pollitt (2004) believes that in principle judging performances directly rather than by indirect scoring should be more valid as the construct that the assessment aims to assess will be the actual criteria that are used when making judgements. The aim of this study was to investigate whether the use of paired comparisons instead of traditional marking could allow the style of questions that could be asked in exams to be less restricted (due to their being no need for reliable scoring) and hence improve the validity of assessment. Although threats to the validity of questions due to the need for reliable marking seem to be most prevalent in newer subjects, geography was chosen for the study in order to be considering a mainstream popular subject and one that tends to contain a variety of very short and slightly longer answer questions in its exams.

Method

Participants

The participants in this study were twenty-seven Year 11 students (aged 15-16 years old) due to take their GCSEs (school leaving exams) later in the year. These students were from three mixed ability classes studying to take Geography GCSE Avery Hill Syllabus. Only those students who sat the higher tier version of their school mock exam, and were

likely to be entered for the higher tier paper[1] in the actual exam, were included in the study. They were attending a secondary school in the south east of England.

Rank-ordering exercise

Of the three classes involved in this study two were taught by one teacher and the other by a second teacher. Around the same time as the students sat their mock exams (January) a researcher visited the school and asked the teachers to carry out an activity of placing the higher tier students in order of geography ability. The teachers were intentionally not pre-warned of this activity so that they would not be able to refer to, for example, past test results when a more holistic judgement of geography ability was desired. Both teachers were provided with small cards printed with the students' names to be shuffled into order; they were asked to arrange the cards into the rank order that best reflected the achievement that the students deserved – which may not be necessarily the same as the order that an exam would place them in.

The teacher who taught two of the geography groups was asked to produce either one rank order of students for both classes if she felt able to do so satisfactorily, or to provide a separate rank order for each class. She started by arranging the cards into two separate orders for the two separate classes and then found she could merge these into one. The two teachers were asked whether they felt they could satisfactorily merge their rank orders together into one overall order but, as anticipated, the teachers did not feel this was possible since neither of them knew all the students sufficiently well.

Modified test paper

A test paper was constructed by the researchers with the assistance of a question writer involved in the Avery Hill Geography GCSE syllabus. The assessment of this GCSE takes the form of two written papers (for higher tier candidates Paper 2: Knowledge, and Paper 4: Decision Making Exercise) as well as a coursework component. The adapted test paper was based on two of the three questions that made up the Year 2000 version of Paper 2. The students had not previously used this paper for practice. The case study element at the end of each question was omitted to keep the test duration within the length of one school lesson. With the assistance of the experienced question writer, the two questions were adapted to be more like they might have been if the demand for reliable marking were removed. The purpose of the study was explained to the examiner in advance; he was happy that the modified test paper was more like it might have been if traditional marking was not going to be used.

Part of one of the original questions is shown below. Pie graphs showing the proportions of primary, secondary and tertiary industry in the UK and in Ghana were presented with this question.

[1] Most GCSE subjects are examined using differentiated papers set at different levels of difficulty. With geography, as in most tiered subjects, the foundation tier covers grades G to C and the higher tier covers grades D to A*.

(ii) **Compare** the employment structure of the UK with that of Ghana. [2]

(iii) **Give reasons** to **suggest** why there are differences between the employment structures of the two countries. [3]

Below is the original mark scheme for these question parts.

(ii) Mark awarded for comparative description of two differences.
e.g. UK has lower primary sector by 44% (1); UK has higher secondary sector by 16% (1) UK has higher tertiary sector by 28% (1) Figures not essential. 1 + 1 = **[2]**

(iii) Explanations require 3 simple statements or one simple plus 1 elaborated. Must refer to at least two sectors
e.g. ***Primary:*** In UK *greater* use of machinery on farms releases people from the land (1); or in Ghana lower stage of development so more labour needed in producing feed (1)
Secondary: In Ghana *fewer* raw materials processed in the country so fewer jobs in secondary (1); or UK has been through industrial revolution so manufacturing industry employs more (1)
Tertiary: Ghana can only afford a *lower* level of services meaning fewer job opportunities (1); or UK reached MEDC stage so technology allows development of more service industries (1)
(1 + 2) or (1 + 1 + 1) = **[3]**

Obviously this question had to be written to be marked reliably on the basis of one mark for one point made. This resulted in a rather rigid expectation of what students would write which is not wholly natural. The lack of need for reliable marking allowed these question parts to be merged into one question, keeping an indication to the students that two different things are required. The modified question is shown below.

(ii) **Compare** the employment structure of the UK with that of Ghana.
Suggest **reasons** for the differences. [5]

Combining the questions allows students to structure their answer more freely. The advantage of this is that examiners can judge the quality of a more open answer, which is likely to improve the validity. One possible disadvantage of such changes is that some students may not be well enough practiced at structuring their own answers and may forget to include things that they know. This problem would be likely to diminish if the method were used in real examining.

Mark allocations for each question were included on the paper as normal to give students an indication of the amount to write for a good answer. The modified test paper was notionally out of 44 marks compared to the 90 marks usually available and was one hour

in length compared to the usual one hour and thirty minutes, or two hours in years previous to 2003. A copy of the modified test paper can be found in appendix A.

Testing

Approximately three weeks after their school mock exam the students attempted the modified test paper under test conditions. This time delay was necessary to avoid increasing pressure on the students at the time of their mock exams. Their teachers warned students in advance that they would be sitting a test for the exams syndicate which would provide them with additional exam practice. This action of emphasising the importance of the test was instigated by the teachers not the researchers but was almost certainly beneficial in encouraging the students to apply themselves to the test in a serious way, more closely replicating the conditions of their mock exam. The one-hour time limit gave sufficient time for most students and few were still writing at the end.

In order to provide the school with some rapid feedback the adjusted test scripts were traditionally marked by a researcher according to the original mark scheme. Due to changes to the papers this marking of the scripts is likely to have been less reliable than would normally be acceptable in exams (not least because it was not carried out by a trained examiner) but its purpose was to provide some initial feedback to the teachers on how the students handled the questions and as such this data can be deemed useful.

Paired comparison exercise

Four experienced examiners involved in the Avery Hill Geography GCSE syllabus attended a two-day exercise to make comparisons between the quality of the scripts from the adapted test.

Students' names were removed from the scripts and scripts were labelled with letter codes instead. Nothing was written on the scripts during the 'traditional marking' so they were blank apart from the students' responses. The rationale for the project was explained to the examiners and they were given specific instructions to do with the comparison exercise. They were asked to compare two scripts at a time and to record which of them represented the better performance. The judgement was to be a holistic judgement. The examiners were asked not to consider the marks that might normally have been awarded, but to read the script through and decide which script they felt represented a pupil who was better at Geography. The examiners were each given script lists showing the comparisons they were to make. These were prepared in advance in order to ensure that all the required comparisons were made and to ensure no unnecessary duplication.

The script lists were arranged to encourage 'chaining' most of the time i.e. after each comparison one script was returned to the central pile whilst the other was retained to compare with a 'new' script. This reduces the reading time involved in making paired comparisons, and no bias effect has been detected in any of the comparability studies that have used chaining procedures.

Prior to this exercise the rank orders provided by the two teachers were merged into one, using the only other information that was available at that time – the marks on the modified test paper. To do this, a regression equation was calculated for each teacher's group of students using statistical software. Each regression equation predicted the rank position given by the teacher based on the mark on the modified test.

$\text{Rank}_{\mathbf{P}} = \mathbf{a} \times \text{Mark} + \mathbf{b}$

$\text{Rank}_{\mathbf{S}} = \mathbf{c} \times \text{Mark} + \mathbf{d}$ where the subscripts **P** and **S** indicate the two groups, and **a**, **b**, **c** and **d** are the constants in the regression equations. The ranks for the smaller group, **S**, were then converted onto the same scale as group **P** by rearranging the equations to eliminate the 'Mark'.

$$\text{Rank}_{\mathbf{S}\text{ on }\mathbf{P}\text{ scale}} = \mathbf{a}/\mathbf{c}\ \text{Rank}_{\mathbf{S}} + (\mathbf{b} - \mathbf{ad}/\mathbf{c})$$

This provided ranks for the students taught by teacher **S** on the scale constructed for those taught by teacher **P**. For example, a student ranked 2nd by teacher **S** received an estimated rank of 4.33 on the scale for group **P**, hence placing him/her between the students ranked fourth and fifth by teacher **P**. This allowed all 27 students to be put into a single rank order. When this was done there was one obvious outlier that was affecting the outcome – a student who had clearly not taken the exercise seriously. This student was removed from the regression analysis and the process was repeated to create the combined rank ordering.

A problem can arise with regression rescaling if the bias towards the mean is different for the two groups. However, the size of this effect depends wholly on the correlation between rank and score within the groups, and here the correlations were identical to two decimal places ($r_{\mathbf{S}} = 0.843$, $r_{\mathbf{P}} = 0.836$) indicating that there is unlikely to be a problem.

To reduce the number of comparisons that the examiners needed to make during the paired comparison exercise the combined rank order was used to eliminate comparisons between the very best students' scripts and those of the weakest. Each comparison that was included was made by two of the four examiners. The pairs to be compared were arranged in the lists so that a pair of scripts compared early on the first day by one examiner were not compared with each other by another examiner until much later to minimise effects associated with order and acclimatisation to the task.

Mock exam marks

The school's mock papers consisted of questions from past exam papers which the students had not previously seen and were put together and marked by the teachers using the original mark schemes. The students' marks for their mock exam were later obtained from the school. One student did not sit one of the mock exam papers and was therefore omitted from the analysis.

Analysis and Results

Rasch analysis of the judgement data generated estimates of the 'ability' of each student (or the 'quality' of each script), with standard errors and misfit statistics. There was no significant misfit of any script or judge in the data set.

A rank order of the students was then created for each of the data sets: – the paired comparison parameter estimates, the scores on the mock knowledge paper, the percentage score on the complete school mock exam and the traditional marking of the modified paper.

As already mentioned, the rank orders created by the two teachers were merged into one rank order using the scores from the traditional marking of the modified test paper. Correlations between the various rank orders are shown in table 1 below.

Table 1 – Correlations between rank orders

Correlations:	Teacher rank order combined using mark on modified paper	Rank from paired comparison exercise on modified test scripts
Rank by mark on mock knowledge paper	0.747	0.833
Rank by overall percentage on mock	0.798	0.853
Rank by mark on modified test	0.821	0.863
Rank from paired comparison exercise on modified test	0.818	-

All the correlations are quite high, for example, the paired comparison rank correlated well with the rank by overall percentage on the school mock (r = 0.853). This indicates that the paired comparison exercise produced an order that correlates well with other more traditional measures of student ability, and that 'paired comparisons' is a valid approach to assessment.

The combined teacher ranking correlates slightly better with the ranking by paired comparison data (r = 0.818) than with the overall mock percentage score (r = 0.798). If we were to assume that the teachers know the students best (as some countries do even for high stakes purposes) then this indicates that the order created by making holistic judgements about the students' work on the modified paper is a slightly more valid measure of the students' abilities in geography than adding up their scores on questions from past papers. Table 1 also shows that the correlation of the combined teacher rank order with the rank on the mock knowledge paper was lower (r = 0.747) than its correlation with the rank order by traditional marking of the modified paper (r = 0.821). This indicates that modifying questions to how they might have been if reliable marking was not a restriction led to an improvement in the validity of the assessment. This is reflected again if we compare the paired comparison rank order's correlation with the rank by mock knowledge paper (r = 0.833) and its correlation with rank by traditional marking of the modified paper (r = 0.863).

However, the combined teacher ranking used here was created using the traditional marking of the modified paper. The modified paper was not designed to be marked in this way and is likely to correlate fairly well with the paired comparison ranks and modified paper traditional mark rank since they all rely on the same evidence. For these reasons, the combined teacher ranking is not a perfect merging of the teachers' rankings and some caution should be taken in considering the correlation between the combined teacher ranking and the other rank orders.

Because of this, the rank orders provided by the teachers were merged twice more, once using the students' overall percentage scores on their school mock exam, and once using

the parameter estimates provided by the analysis of the judgements from the paired comparison exercise. This was done in a similar way to before but using either the percentage score on the overall school mock exam or the parameter estimates from the paired comparison exercise as the common scale. Table 2 below shows the correlations of these two merged teacher rankings with the rankings from the mock test, paired comparison exercise and traditional marking of the modified paper.

Table 2 – Correlations between the combined teacher rankings created by merging using the paired comparison evidence and the mock percentage score with the other rank orders.

Correlations:	Teacher rank order combined using overall percentage on mock	Teacher rank order combined using paired comparison data
Rank by mark on mock knowledge paper	0.792	0.786
Rank by overall percentage on mock	0.834	0.841
Rank by mark on modified test	0.829	0.833
Rank from paired comparison exercise on modified test	0.841	0.855

Table 2 shows that the combined teacher ranking based on the paired comparison data, like the first combined ranking, correlates slightly more highly with the paired comparison ranking ($r = 0.841$) than with the rank by overall school mock percentage score ($r = 0.834$). Again this indicates that the former is providing a good measure of ability and is probably a slightly more valid measure. Again, the combined teacher rank correlates more highly with the rank by modified paper traditionally marked ($r = 0.829$) than the rank by score on the knowledge mock paper ($r = 0.792$) again indicating that the modified paper is probably more valid. The same pattern is also seen for the teacher ranking combined using the overall percentage scores on the mock.

None of these three common scales for combining the teacher rank orders is perfect but the fact that using any of them shows the same pattern is an indicator that the pattern is genuine and that the modifications made the paper more valid and the paired comparison judgements made a more valid ordering of ability than the mock marks.

Discussion

In this study we have taken the teachers' rank ordering of students 'quality' as a criterion measure. There is a long history of using teachers' rankings in this way in test validity studies in primary schools (e.g. Southgate, 1962; Gillham and Hesse, 1970; France, 1979; Young, 1980; Pollitt and Abramovici, 1986), but less in secondary schools. On the other hand, many countries use teachers' ratings as a significant, or even the only, component

of a school leaving certificate. In this study both teachers had known their pupils for some time and assured us they knew them well enough; they took only a few minutes to sort the name cards and reported no problems in doing so.

Assuming that the teachers know the students best, the correlations between rank orders suggest that the paired comparison method using scripts from the modified paper provided more valid results than the school's mock exam which was marked traditionally. In addition, the modified paper marked in a traditional way provided a more valid assessment of students than the equivalent traditional past paper used in the mock exam.

The differences between the validity coefficients were small, as expected. Current examination procedures are highly valid, and there was therefore not much room for improvement. Our modified paper involved only two questions, and was about half the size of a normal paper. The analysis was complicated by the need to merge two teachers' rankings. Yet the results were in the desired direction, whichever merging method we used: re-designing the questions to facilitate judgement increased the validity, and using paired comparison methodology increased it further. Of course our sample, too, was small with just 26 pupils in the final analysis, and we must therefore not claim that the findings would generalise with too much confidence.

The experience of the study raised two particular issues for consideration if the method were to be used in real examining. Firstly, the examiners reported not feeling entirely confident about the judgements they were making and about whether they were making their judgements consistently. Such feelings seemed to reduce a little as the exercise went on but did not go away. One concern raised was that they might not all be placing the same value on the same kinds of skills or qualities (e.g. factual accuracy as opposed to application of knowledge or explanation of a point). The examiners felt that an initial discussion of the kinds of answers to each question that should be considered good or weak and of the amount of emphasis to be placed on certain skills (at certain levels) would have helped. They suggested that a set of written judging criteria or guidelines would have made them feel more confident that they were making consistent decisions – in other words, they needed a replacement for the traditional marking scheme.

Of course, in an operational system the misfit analyses will identify any examiner who deviates far from the consensus, and in this study no one did, but the issue of proper preparation so that examiners feel confident from the start still needs to be addressed.

Another practical issue that arose was how to deal with a rubric infringement (the omission of a question). Even given the guidance provided, the examiners seemed to feel a little uncomfortable making judgements about this script but this may have been because omitting one question of the two left only a limited amount of work from which to judge. Guidelines to examiners would be needed to explain how to deal with rubric errors.

There are a number of other concerns about the paired comparison method that would need to be resolved in order for the method to work. For example, whether the system would be publicly accepted generally, whether using teachers' predicted grades as an initial filter for the comparisons to be made could be accepted, and whether the method

would be understood or could lead to accusations of unfairness. See Pollitt (2004) for a full discussion of these concerns.

In conclusion, it seems that judgements made by several examiners rather than just one or two should reduce risks of unfairness. The fit statistics that would be available should also reduce risks of unfairness by prompting the collection of additional judgements about problematic scripts or scripts that are close to a grade boundary.

This study indicates that the paired comparison method could lead to more valid assessment by reducing the restrictions placed on the way that questions are written when traditional marking is to be used. The method offers lots of potential advantages, and as such warrants further investigation.

References

Ahmed, A and Pollitt, A (2000) *Observing Context in Action*, International Association for Educational Assessment Conference, Jerusalem, May 2000, available at http://www.ucles-red.ac.uk/conferencepapers.htm

France, N (1979) *The Primary Reading Test*. Windsor, NFER-Nelson.

Gillham, WEC and Hesse, KA (1970) *The Leicester Number Tes*t. Sevenoaks, Hodder & Stoughton.

Good, FJ and Cresswell, MJ (1988) Can teachers enter candidates appropriately for examinations involving differentiated papers? *Educational Studies*, 14, 3, pp 289-297

Hambleton, RK and Swaminathan, H (1985) *Item response theory: Principles and applications*, Boston, Kluwer-Nijhoff

Murphy, RJL (1981) Symposium: Examinations: O-level grades and teachers' estimates as predictors of the A-level results of UCCA applicants, *British Journal of Educational Psychology*, 51, pp 1-9

Pollitt, A (2004) *Let's stop marking exams*, International Association of Educational Assessment Conference, Philadelphia, June 2004, available at http://www.ucles-red.cam.ac.uk/conferencepapers.htm

Pollitt, A and Abramovici, S. (1986) *Wordsearch*. Sevenoaks, Hodder & Stoughton.

Pollitt, A and Elliott, G (2003) *Finding a proper role for human judgement in the examination system*, Qualifications and Curriculum Authority Seminar on 'Standards and Comparability', April 2003, available at http://www.ucles-red.cam.ac.uk/conferencepapers.htm

Pullinger, H (1982) Marks, internal assessments and external examinations, *Teaching Geography*, 8, 1, pp 22-23

Southgate, V (1959) *Group Reading Tests: Test 1*. Sevenoaks, Hodder & Stoughton.

Thurstone, LL (1927) A law of comparative judgement, *Psychological Review*, 34, 273-286

Wood, R (1991) *Assessment and Testing: A Survey of Research*, Cambridge, Cambridge University Press

Young, D (1980) *Group Reading Test*. Sevenoaks, Hodder & Stoughton.

Appendix A – Modified Test Paper

UNIVERSITY OF CAMBRIDGE LOCAL EXAMINATIONS SYNDICATE

General Certificate of Secondary Education

GEOGRAPHY SPECIFICATION B (AVERY HILL) **1987**

Additional materials:

Resource Sheet (1987/RS) – inserted

16 page answer booklet

TIME 1 hour

INSTRUCTIONS TO CANDIDATES

Write your name in the space at the top of the separate Answer Paper.

This question paper contains two questions. Answer **both** questions.

Answer **all** parts of the questions in your Answer Booklet. Make sure each answer is clearly numbered.

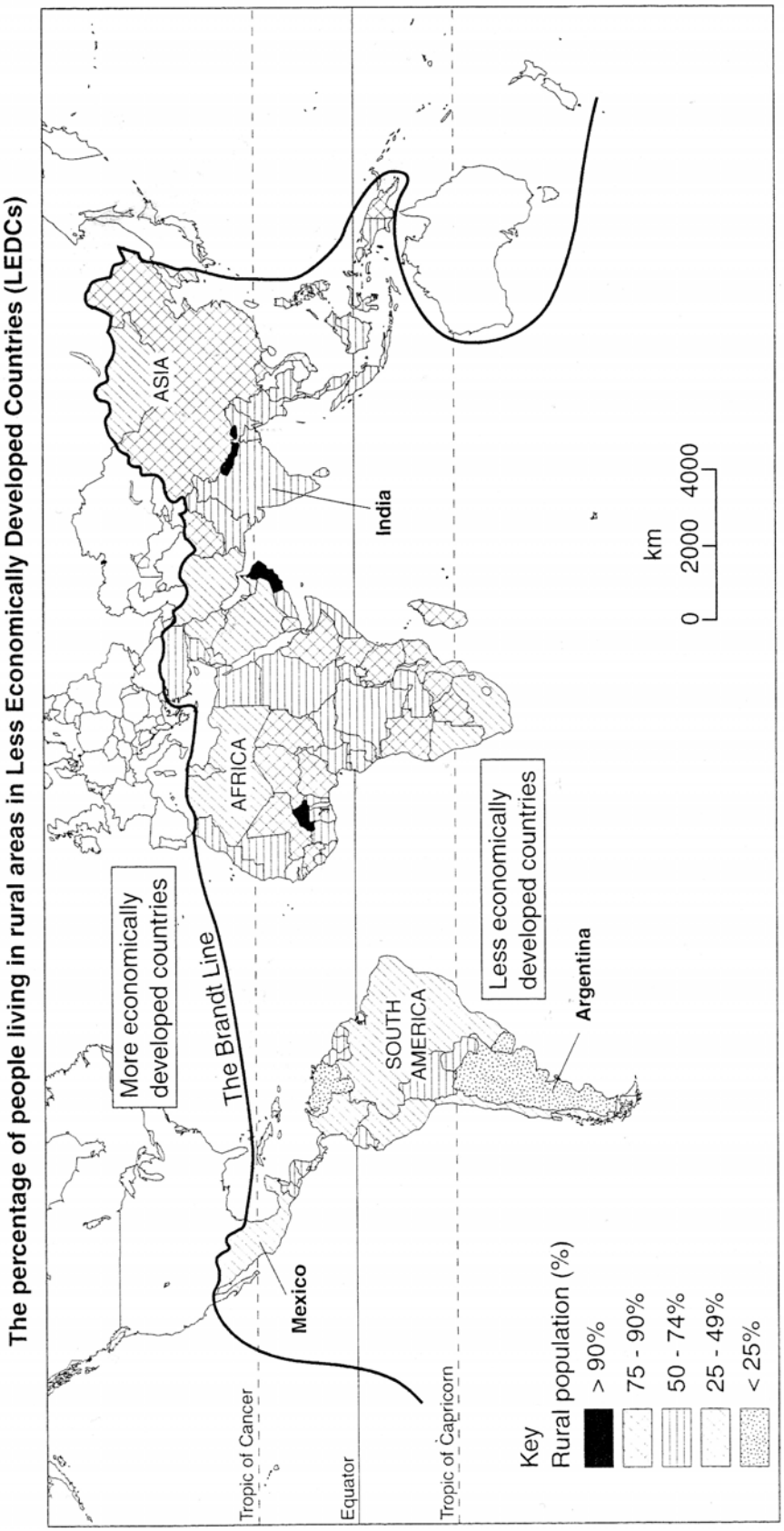
INFORMATION FOR CANDIDATES

When answering the questions in this test, use the number of marks given in brackets [ ] at the end of each question or part question as an indication of how much to write.

There are **44 marks** available on this paper.

PEOPLE AND PLACE

Question 1



(a) **Study** the map opposite

(i) **Compare the** percentages of the population living in rural areas in India, Argentina and Mexico. [3]

(ii) **Describe** the **distribution** of areas where the rural population is more than 75% [3]

(b) **Study** MAP 1 on the **Resource Sheet**. It shows the village of Aminbhavi near to Mumbai (Bombay) in India.

(i) **Describe** the **distribution** of the homes of different groups of people and **explain** how these groups benefit from the local services. [4]

(ii) Push and pull factors cause people to move out of villages to live in cities in LEDCs.

Suggest and explain reasons why some villagers are moving from Aminbhavi into Mumbai. **Refer to push factors only.** [4]

(iii) **Explain** the problems caused for planners and city authorities when large numbers of people move into LEDC cities. [8]

Total mark 22

End of Question 1

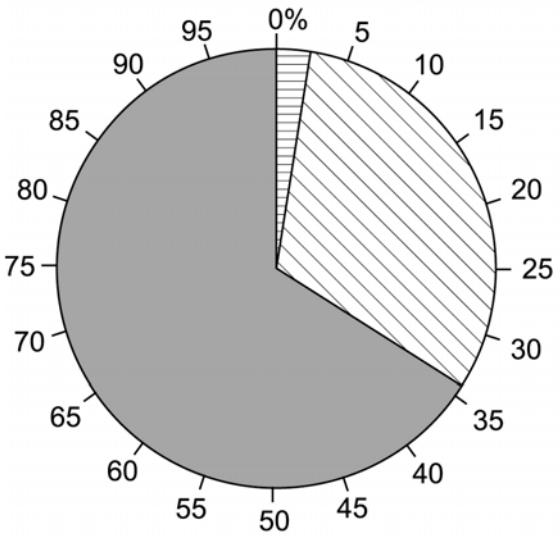
BLANK PAGE

PEOPLE, WORK AND DEVELOPMENT

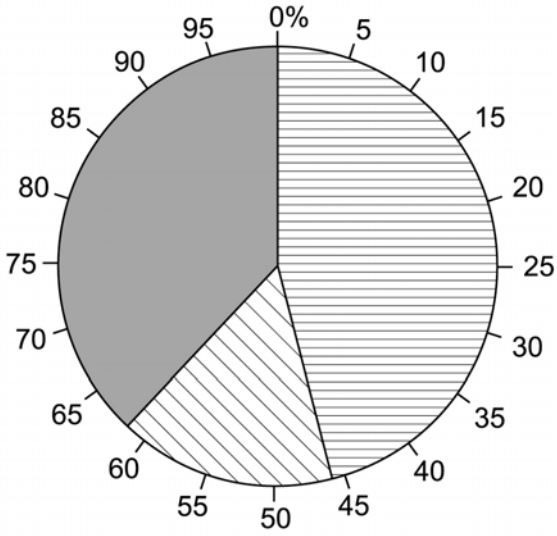
Question 2

(a) **Look at** the pie charts below. They show the employment structure of the United Kingdom (a More Economically Developed Country) and Ghana (a Less Economically Developed Country) in West Africa.

Employment structure in the United Kingdom (UK)



Employment structure in Ghana



Key

Primary industry

Secondary industry

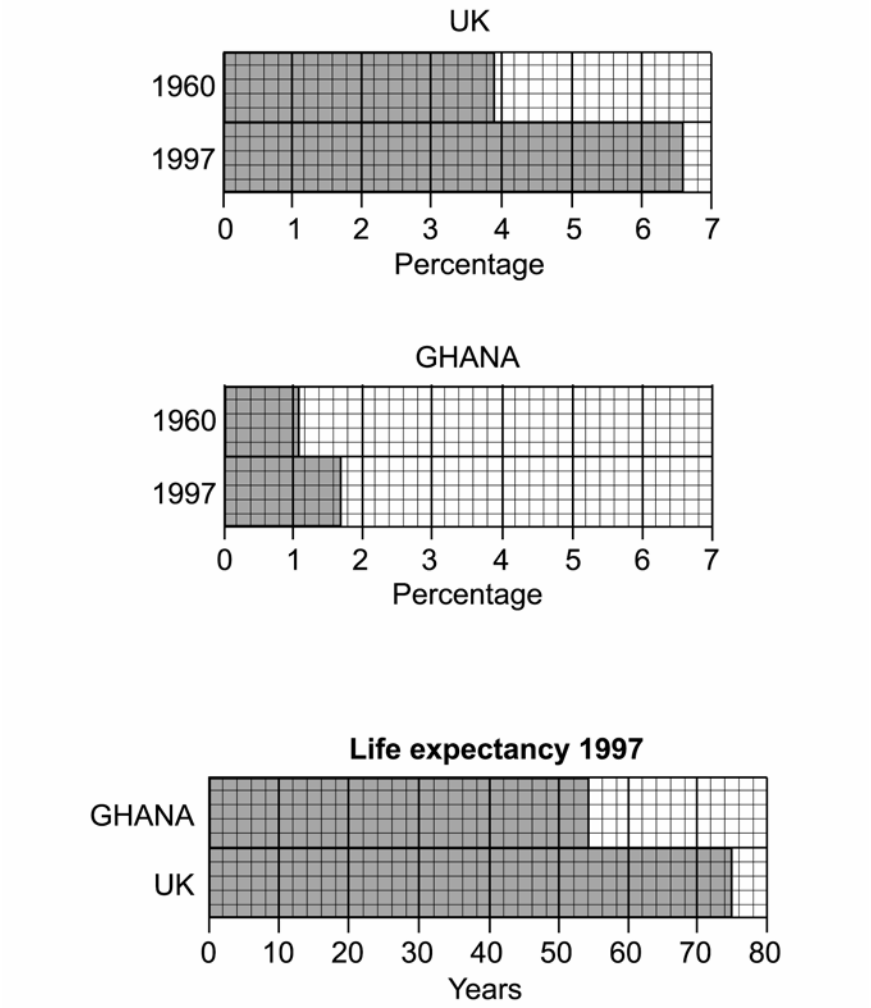
Tertiary industry

(i) What is meant by each of primary, secondary and tertiary industry? [3]

(ii) **Compare** the employment structure of the UK with that of Ghana.
Suggest **reasons** for the differences. [5]

(b) **Look at** the two graphs below.

The percentage of each country's money spent on health care



(i) **Compare** changes in spending on health care in Ghana and the UK between 1960 and 1997. [2]

(ii) **Suggest** how this **and** other factors may affect life expectancy in the two countries. [6]

(c) **Read** the news report below. It is about the Volta Dam in Ghana.

Ghana's Volta scheme - successful or not?

Modern industry was attracted by the cheap electricity produced at the Volta Dam. The new industry mainly used machinery but provided jobs for some people. Many men migrated to the area in the hope of getting a job, leaving their families behind in villages. In a country where many people are not in paid employment, creating labour-intensive jobs in villages may have been a better use of the money than the capital-intensive Volta scheme.

(i) Do you think that Ghana should have used a capital-intensive large-scale project like the Volta scheme or would a labour-intensive small-scale industry have been better? Justify your choice. (Refer to other places you have studied in your answer if you wish.) [6]

Total mark 22

End of Question 2

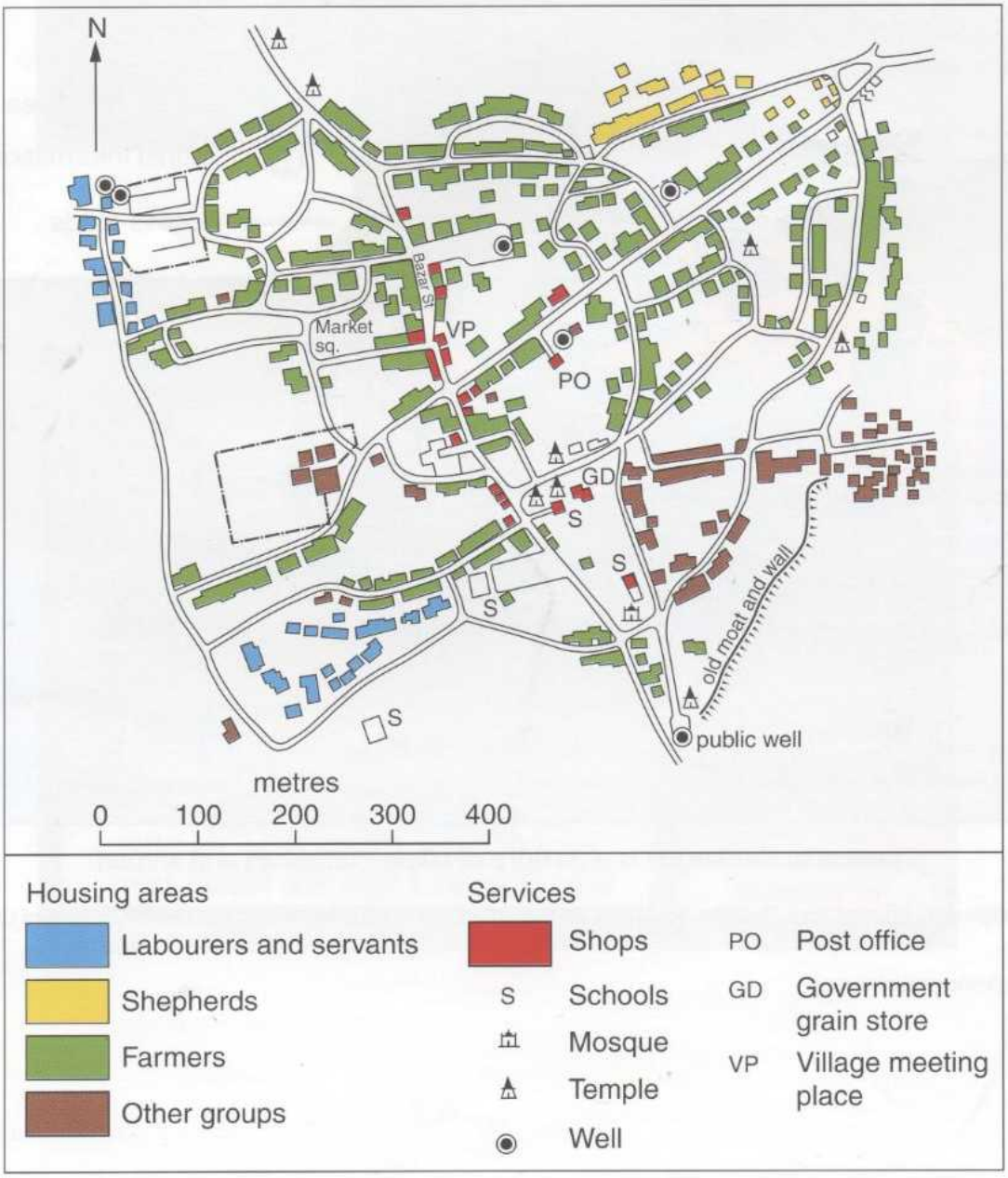
General Certificate of Secondary Education

University of Cambridge Local Examinations Syndicate

GEOGRAPHY SYLLABUS B (Avery Hill) 1987/RS

RESOURCE SHEET

Map 1 for use with Question 1



Aminbhavi is a small village near Mumbai (Bombay) in India.

Acknowledgement: Adapted from 'Poverty and Wealth in Cities and Villages', Simons (Oxford University Press)