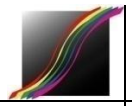
Educational Measurement and Evaluation Review (2010), Vol. 1, 116-139



http://pemea.club.officelive.com/ EMEReview.aspx

Determining the degree of inter-rater consistency in a high stakes, developing country, examination

Jack Holbrook
Visiting Professor in Science Education, University of Tartu, Estonia

Abstract This study examines the reliability of scoring (awarding marks) for essay-type questions in a high stakes, external examinations where marking schemes are not well developed. The study compares the inter-rater reliability between the initial marks awarded in an actual examination with a first remarking where no additional guidance was given and with a second remarking following a two day training programme. As with studies in other countries, the findings show poor correlations between marks awarded and substantial changing of grades on remarking. The additional guidance given in the workshop was found to increase reliability as shown by the correlation with a purposive remarked sample by an expert from the examination board, but the potential to use double marking as a mean of increasing reliability was shown not to be appropriate. Suggestions are made to develop question setter-produced marking schemes and the use of sample marking to give greater guidance to markers to raise the reliability of results.

Keywords: *Inter-rater consistency, high stakes, examinations, reliability, Rasch analysis*

Introduction

The setting of examinations in various subjects by an examination board external to the school is common around the world. The examinations are taken at the end of a course, or school curriculum, and determine success or failure for future advancement and are often referred to as once only, high stakes, examinations. These can involve objective testing, often multiple choice (MCQ) items, and/or more subjective questions where students construct the responses themselves. Where MCQ items are involved, the scoring is dichotomous, undertaken by machine and reliability is less of a concern than validity of the items chosen. As it is generally accepted that valid instruments are important, essay-type questions still remain an obvious choice for evaluation of knowledge (Verma,

Chhatwal, & Singh, 1997). Unfortunately allocating marks for essay-type, student free-response questions can be very unreliable, unless marking systems are well developed (Baird, Greatorex, & Bell, 2004). In fact, poor intra- and inter-rater reliability of rater evaluations in free-response assessments have long been recognized (Van der Vleuten, 1996). With this in mind, it has been suggested that awarding bodies take clear steps to ensure examinations are marked reliably and procedures are detailed in a code of practice (Baird, Greatorex, & Bell, 2004; Newton, 1996).

Reliability is usually taken to mean the reproducability of scores on another occasion. Reliability has been defined (Berkowitz, Wolkowitz, Firch, & Kopriva, 2000) as the degree to which test scores for a group of test takers are consistent over repeated applications of a measurement procedure and hence are inferred to be dependable and repeatable for an individual test taker. An index of reliability of a test for a given population can be taken to be the ratio of true score variance to observed score variance (Dimitrov, 2002). True scores cannot be directly determined and hence the reliability is an estimate. And of course, this assumes that the scores were obtained from a test that was seen as sufficiently valid for the reliability of the results to have any meaning at all. A very reliable mathematics test would be totally out-of-place as a reliably instrument for the determination of history ability! But if validity is at an acceptable and interpretable level, then reliability is an important consideration facing all examination boards (Brooks, 2004).

Reliability of test scores can be influenced by the type of performance data and the metric in which the scores are express (Yen & Candell, 1991). Unreliability in the assessment system can be traced to a number of factors, such as:

(a) Use of short tests
(b) Use of essay scripts where the marking is subjective
(c) Lack of rater guidelines
(d) Rater variability
(e) Bias in the distribution of examination scripts
(f) Lack of structuring of essay-type questions

(a) Use of Short Tests

In such tests there are insufficient items to sample the domain of learning and hence leads to unreliable measures. Steps to overcome this can, of course, be to increase the length of the examination, but then there is the question of examinee fatique and if the test is over lengthy, the score again becomes unreliable. Another approach is to make use of MCQ items which are answered quickly, marked reliably and allow coverage of the curriculum content. The MCQ items can then be coupled with a shorter array of other type of questions which sample the testing domain.

(b) Use of Essay Scripts Where the Marking is Subjective

Impression marking has been shown to be unreliable without carefully developed guidelines (Baird, Greatorex, & Bell, 2004). Verma, Chhatwal, and Singh (1997) claimed that the 'structuring' of essay questions provides a tool for improving

rater reliability. In their study, they concluded that structuring an essay question helps improving rater reliability as well as the internal consistency of the test, without any additional input. This can be an important consideration for testing subjects where creativity, or the testing of free writing ability, is not an important issue. It can be seriously considered for science and social science subjects.

(c) Lack of Rater Guidelines

Reliability can be expected to be problematic if there are no marking guidelines, or the guidelines are inadequate. A range of strategies has been adopted to increase the reliability of marking. Among these are better specification of scoring criteria, including in some cases the use of analytic (awarding marks and even half marks, for each specific component) rather than holistic rating scales (indicating a mark range and expecting the rater to indicate whether the script being marked is high medium or low within the range) (Baird, Greatorex, & Bell, 2004; Elder, Knoch, Barkhuizen, & Von Randow, 2005; Moon & Hughes, 2002; Van der Vleuten, 1996;). The goal is to move away from impression marking.

(d) Rater Variability

Variability in rater behaviour may produce invalid, or unfair results for candidates whose absolute scores or relative positions may differ depending on who assesses their performance. To address this, rater training has been suggested (Baird, Greatorex, & Bell, 2004: Elder, et al., 2005; Munro, et al., 2005; Weigle, 1998). The training usually consists of a number of raters being (re)introduced to the scoring criteria and then asked to rate a number of scripts as a group (sample marking). Within the training, ratings are carried out individually, then discussed by the whole group and reasons for discrepancy clarified. The discussion rarely covers the awarding of full marks as this is usually agreed, but the focus of the discussion relates to the manner in which partial marks are awarded and the degree to which explanations given are to be considered appropriate. Where a candidate has written a weak response, clearly far less than adequate, much discussion often takes places as to whether this should be awarded any marks at all.

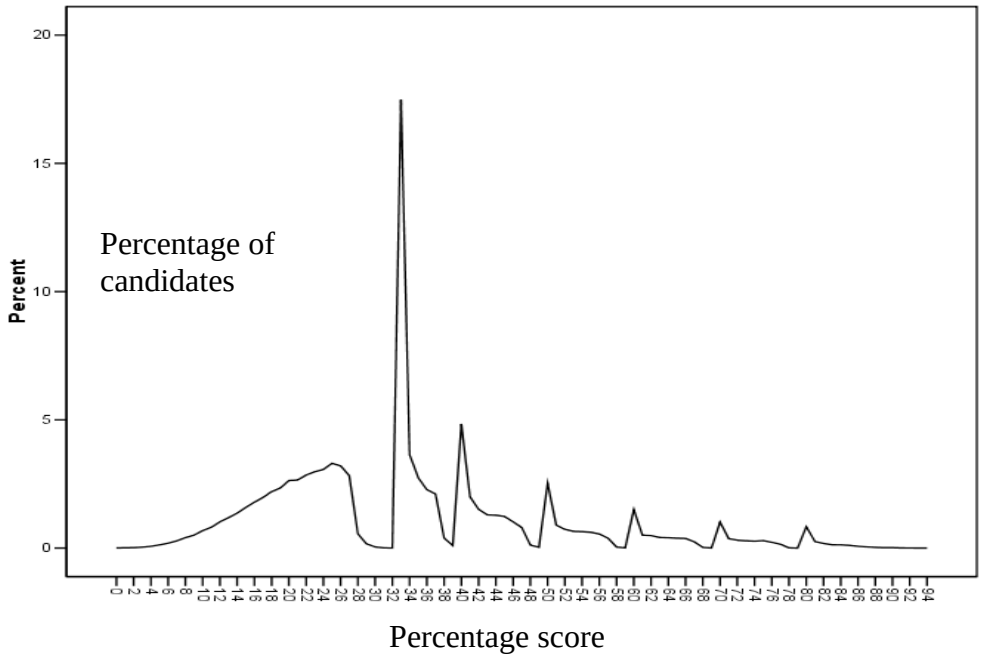
Some of these scripts used for training purposes can be specifically chosen because they correspond to coverage of different score levels on the rating scale, or they might exemplify certain problematic, or unusual issues arising in written assessment. Rater training has been shown to increase the self-consistency of individual raters by reducing random error, to reduce extreme differences between raters, to clarify understanding of the rating criteria, and to modify rater expectations in terms of both the characteristics of the writers and the demands of the writing tasks (Weigle, 1998), although the effectiveness of such training does not make raters 'duplicates of each other' and increases rater consistency rather than rater severity of marking. Unfortunately Lunz, Wright, and Linacre,(1990) note that raters (or judges) often employ unique perceptions which are not easily altered by training. Such perceptions could be to finish the marking as quickly as possible, or to ensure that candidates near the borderline are not failed on account of any rater.

(e) Bias in the Distribution of Examination Scripts

Bias can occur if examination scripts are not randomised and one rater gets a collection of scripts which are all worthy of high marks, while another rater obtains low scoring scripts only. This can lead to a tendency to be lenient with the low scoring candidates and more harsh with the high scoring candidates. Randomisation can be expected to lead to a range of scores for the scripts marked and hence allow the marking to have a more full perception of the marking range.

This stems from situations were little control is exercised over the marking process and raters are aware that the marks they award will affect pass rates. It can lead to situations where raters award 'grace' marks (extra marks to enable the candidate to reach the pass mark or to reach a mark that would enable a change from one grade score to another). Figure 1 gives a typical outcome taken from a SSC examination in Bangladesh. The pass mark in this examination is 33% and grade awarded to candidates change at 40%, 50%, 60%, 70%, and 80%.

Figure 1
Percentage Distribution for Scores in One Examination by one Examination Board in Bangladesh



(f) Lack of Structuring of Essay-Type Questions

A study, undertaken in India (Verma, Chhatwal, & Singh, 1997), compared the effect of structured essay type questions on rater reliability from a large-scale public examination in history. The study clearly brought out the low reliability of traditional essay questions (TEQ). Further objective evidence of this fact was provided by analysis of the variance of mean marks awarded by seven examiners to the entire group. The internal consistency of the TEQ paper was found to be poor, although a structured essay paper (SEQ) gave a significant internal consistency. Unfortunately they suggested that providing a model answer or check lists to the examiners, are not feasible, at least in their circumstances, because the person setting the paper is usually different from the person evaluating it and thus the chances of agreeing to a common check list are lower. Furthermore, with large numbers of candidates marking TEQs, the time-consuming marking process does not accommodate marking by more than one rater.

Verma, Chhatwal, & Singh (1997) thus claimed that the 'structuring' of essay questions provides an important tool for improving rater reliability. In fact, they concluded that structuring essay questions helped in improving rater reliability as well as the internal consistency of the test, without any additional input. Making the marking of questions less dependent on the rater is certainly an important consideration in subjects where free writing is not considered essential for construct validity.

Reliability of Raters

It has long been recognized that the scoring of essay type questions is a complex process, with many variables, in multiple combinations, influencing the reader differentially (Chase, 1986). Single rater reliabilities (correlation between responses by two raters) can vary between .3 and .8 depending on 'the length and topic of the essay, the amount of freedom students have in selecting and responding to the essay, the experience of the raters, the extent of training given the raters and the control exercised in the marking environment' (Marsh & Ireland, 1987).

Inter-rater reliability refers to reliability of marking betweeen raters Unfortunately raters vary in their degree of leniency or harshness and the average proportion of variance due to rater effects can be as high as 49% (Eckes, 2005). To address this a number of researchers have suggested rater training (Baird, Greatorex, & Bell, 2004; Weigle, 1998). A faily simple training approach is to introduce sample marking in which each rater marks the same scripts and then their individual marks are compared, discussed and adjustments made to the marking scheme so as to further clarify the expected marking procedure as put forward. If this is carried out with scripts across the mark range so that raters can agree on the meaning of full marks and partial marks at a number of levels, then the inter-rater marking can be increased (Bédard, Martin, Kruger, & Brazil, 2000). Retraining has been the dominant method to induce judges to make similar assessments of candidate performances (Lunz, Wright, & Linacre, 1990). However its effectiveness is questioned and correcting measure are felt necessary to ensure judge severity is considered more manageable and more fair.

A further important consideration which gained much attention in the mid-20th century is double marking. This involves two raters independently marking each script and the marks compared. It is still a common University practice (Dahl, 2006) and in a country like Bangladesh forms the major attempt at reliability at this level. The public Universities in Bangladesh make use of double marking by sending candidate scripts to two raters and the average of the two marks is taken as the true mark. If the difference between the two percentage marks is greater than 20, then the script is sent to a third rater and the third mark is considered as representative. Johnson, et al. (2005) refer to this as the expert judgement model as the mark replaces boh marks for the other examiners. An alternative is to follow what Johnson, et al. (2005) refer to as the tertium quid model in which the mark is averaged with that of the rater closest to the mark given by the third rater and hence essentially eliminating the rater furthest from the average. Unfortunately the use of double marking is time-consuming and costly and has lost favour at secondary levels where examination boards are under pressure to complete their marking as quickly as possible and also to maintain reasonable costings which do not cover the notion of double marking for large numbers of candidates (Newton, 1996). This study examines the effect of double marking and its effectiveness at secondary level, noting the time consuming factor (Verma, Chhatwal, & Singh, 1997).

Measurement of Rater Reliability

It is often believed that both inter- and intra-rater reliability must be documented to support the adequacy of an instrument. However, inter-rater reliability is unlikely to be stronger than intra-rater reliability, because measurement error is more likely to occur with different raters than with the same rater. Thus, one can be convinced of the instrument's intra-rater reliability if inter-rater reliability is adequate, whereas the opposite is not necessarily true. Nevertheless, it is noted that with fewer raters, it is easier to control the reliability factor (Campbell, 2005).

For instruments using continuous scales, reliability is generally measured with correlation coefficients, or paired *t* tests (Bédard, et al., 2000). Researchers often use them interchangeably in the belief that they produce similar results when applied to assess reliability. Bédard, et al., (2000) recognise this assumption is not correct and gives examples to illustrate the point. The Pearson product moment correlation assesses linear relationships and is not affected by systematic differences between sets of scores, as long as the linear relationship between scores remains the same. Specifically, if one rater consistently rates participants higher or lower than another rater, the resulting correlation coefficient will be unaffected by this discrepancy. Against this, the *t* test detects systematic differences between scores, but not random differences above and below the mean, because it is devised to compare means (Altman & Bland, 1983). Newton (1996) examined the reliability using two indices – the coefficient of correlation and the mean mark differences between the initial prime mark and the re-mark. This gives an indication of the change of rater severity in awarding marks. He found extremely high Pearson product-moment correlations, especially for mathematics, and utilised ANOVA to determine the significance in differences of mean scores. In this study correlation

are employed to show relationships between grades for remarking. The t-test is used to show significance of changes of grades on remarking.

Measurement of internal consistency using Cronbach's alpha, whilst not perhaps as authentic as other methods, is realistic. Although a high coefficient alpha will result when a single general common factor runs through items of a test, high values can be found when item variance is determined by several common factors. In their analysis of GCE examinations in the UK, Nuttall and Willmott (1972) observed that most values lay between .8 and .9 and described this as a 'creditable achievement'. The selected nature of such a student population makes it unlikely that coefficients above .9 could be obtained. In this study, it is preferred to examine inter-rater consistency using Rasch analysis.

Based on analysis of variance, generalizability analysis or theory (Crossley, Davies, Humphris, & Jolly, 2002) accounts for differing methods of reliability estimation as well as different kinds of reliability—between-rater, within-rater, within-candidate, within-teacher. By placing emphasis on estimating variance components rather than effect, the contributions of each source of difference between the observed score and the true scores can be calculated (Brennan & Pediger, 1981). It may then be possible to forecast the number of measurements required to reach a reasonable estimate of the true score. Once key sources of error have been identified, prediction of reliability levels can be made when alternative assessment arrangements are envisaged, e.g., decreasing the number of papers or increasing the number of raters. As the number of papers, marks per paper and the number of scripts marked by a rater are controlled by examination boards, factors associated with changes in these areas were not explored.

Additional Perspectives on Reliability

In previous research, interrater reliability has been studied using Rasch measurement models (e.g., Engelhard, 1994; Eckes, 2005; Elder, et al., 2005; Lunz, et al., 1990; Weigle, 1998). Rasch modelling allows rater severity to be considered alongside the candidate ability and level of difficulty of the essay-type paper. In this study, Rasch analysis is used to examine rater severity and the effects of training on this.

The Bangladesh Practice

How reliable is the marking of essay-type questions in a country such as Bangladesh, where the external examination is of vital importance for a candidate's future education? Little, if any, training is offered to raters and no answers are supplied by the examination board to the questions set. In Bangladesh, at SSC (grade 10) level, an examiner, utilising guidelines from a scheme which gives only a distribution of marks and, prepared by head examiners, plus verbal directives from the controller of examinations, marks a set of candidate answer scripts. These guidelines really amount to little more than the number of marks to be awarded for each question and the degree of leniency to use. As a result, there is a widespread belief that unreliability prevails in marking the SSC scripts i.e. there is a strong

possibility that marking the same scripts by a different examiner would produce very difference results.

As no previous study had been conducted to highlight this issue in the context of Bangladesh, a research study was carried out to determine whether the remarking of SSC scripts would show any significant difference in the marks awarded. As the actual marks are not disclosed to candidates, but a grade used instead, the study considered the change of grade as well as total marks that were obtained on remarking each script. This was possible as no standardisation process is used to match grades to criteria with a view to increasing the fairness of marks awarded.

The study described here was conducted to identify the degree of rater consistency and possible steps that could be taken to improve inter-rater reliability. The paper focusses on inter-rater reliability, in terms of consistency and accuracy of marks given to candidates' written performance on the non-MCQ component of the grade 10 external examinations counting for 50% of total marks in each subject examination.

The study also examined the effect of additional guidance given to raters in terms of training related to a marking scheme per question. This did not involve detailed answers as the questions were sufficiently open –ended to preclude this, but the guidelines were put forward for banding marks, based on descriptors.

In this study the following hypothesis were tested: (1) there is no significant difference in grades, derived from total marks assigned to candidates, on the remarking of scripts; (2) averaging marks from double marking produces a more consistent mark for each candidate; (3) providing limited in-service guidance to examiners produces more consistent marks as shown by a similar distribution in grades per script by raters.

Methodology

The initial marks were taken from those awarded to 4030 candidates in an actual SSC examination. The 1st remarking was undertaken with a mixed group of 11 experienced (9) and new raters (2), marking on a separate occasion, but under similar guidance to that given to the raters during the marking of the original scripts. Each rater marked a separate set of approximately 400 answer scripts. A further remarking (2nd remarking) was carried out after providing the 10 raters with better marking range descriptors and guiding the raters on its use during a 2-day workshop. Each rater marked approximately the same number of randomly assigned answer scripts, a few of which they had marked previously for the 1st remarking. The marks from the second remarking were compared with previous marks.

To carry out the study, an arbitrary script code was recorded on the remaining portion of an OMR (optical machine reading) attachment to the scripts as taken from the earlier SSC examination, as well as to the top of each original answer script. After assigning scripts code, the OMR attachments were removed from the scripts so that total marks from the actual examination were no longer indicated. Any previous rater marks, recorded on the various pages of each answer script, were covered by coloured paper, the scripts randomised and then allocated to a group of

raters comprising both experienced and fresh raters. The raters were not given any guidelines, apart from being supplied with the actual questions used and the number of marks to be assigned for each question; and they were advised to mark the scripts on the basis of their previous experience in internal and external examinations avoiding emotional or sympathetic attitudes i.e. to follow a professional approach. The marks obtained in this way were recorded as the 1st remarking of answer scripts.

After completing the remarking study, the rater were given a short in-service course of two days and guided in the use of a marking approach, developed for this specific purpose by one of the authors. Each rater was guided in the use of this marking scheme by being asked first to mark a sample script. In this, each rater marked the scripts according to the marking guidelines provided and then discussed their marks with the other raters so as to arrive at a consensus on how the marking system is to be interpreted.

Following the short in-service training course, each rater was allotted another set of scripts to mark. No attempt was made to give raters the same scripts as before, although an attempt was made to randomise the scripts so that any one rater did not get all scripts having higher marks, or all the low scoring scripts.

All marks obtained were entered into the computer against the rater number. Marks were totalled and the final mark was record out of 50. As corresponding MCQ marks (which made up the remaining 50% of examination marks) were not available, the marks obtained were converted to grades based on the standard system in use (>40 – A+; 35-39 – A; 30-34 – A-; 25-29 – B; 20-24 – C; 15-19 – D; <15 – F).

To ascertain the expected mark, where marks given to a sample set of 58 scripts differed from each other on the initial, first and second remarking, these scripts were re-marked by a member of the Examination Board familiar with the examination and the subject matter. These marks were compared with the other marks and the correlations determined.

Rasch analysis was conducted using Conquest (Wu, Adams, & Wilson, 1998). The grades given by the raters who participated in the 2-day training programme against the first and second remarking grades were compared to determine inter-rater severity in allocating marks, following a similar study by Weigle (1998).

Findings

The grades obtained from the initial examination, as well as the 1st and 2nd remarking are given in Tables 1, 3, and 5 respectively. These Tables provide a comparison of the grades obtained from the different marking exercises. A correlation of the change in grades between the first set of grades and the grades obtained from the 1st and 2nd remarking is indicated in Table 7, whereas a similar correlation, from the purposive sample additional remarked by an expert, is shown Table 8.

Tables 2, 4, and 6 show the degree of change of grades on remarking from the first, initial marking to the two subsequent sets of remarking. The small type indicates actual changes in numbers of candidates which occur.

Figures 1-4 show the output from the Rasch analysis based on, respectively, remarked scripts allocated to the raters for the 1st remarking, the same scripts re-marked for the second time (largely by different raters), allocated scripts marked by raters for the 2nd remarking, and also the analysis for the scripts allocated to the rater for the 2nd remarking, but referring to the 1st remarking. Figure 5 shows the difference when raters marking the initial, examination scripts are also included.

Table 1
Comparison of Grades Obtained from the Initial Marking and the 1st Remarking

	Grades from 1st remarking															
Grade	A+		A		A-		B		C		D		F		Total	
Initial marking	N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%
A+			1	*33.3*	1	*33.3*	1	*33.3*							3	*0.1*
A	2	*4.9*	10	*24.4*	17	*41.5*	9	*22.0*	3	*7.3*					41	*1.0*
A-	2	*1.1*	17	*9.4*	41	*22.8*	71	*39.4*	33	*18.3*	12	*6.7*	4	*2.2*	180	*4.5*
B			5	*1.4*	42	*11.7*	114	*31.8*	134	*37.6*	49	*13.6*	15	*4.2*	359	*8.9*
C			2	*0.2*	16	*1.9*	113	*13.3*	298	*35.0*	274	*32.2*	149	*17.5*	852	*21.1*
D					2	*0.1*	30	*2.1*	169	*11.7*	510	*34.2*	739	*51.9*	1450	*36.0*
F			-				1	*0.1*	5	*0.4*	59	*5.2*	1080	*94.3*	1145	*28.4*
Total	4	*0.1*	35	*0.9*	119	*3.0*	339	*8.4*	642	*15.9*	904	*22.4*	1987	*49.3*	4030	*100*

Note. *N* = number of candidates

Table 2
Change of Grades from the Initial Marking to the 1st Remarking

	Grades from 1st remarking															
Grade	A+		A		A-		B		C		D		F		Total	
Initial marking	N		N		N		N		N		N		N		N	%
A+	0		-1 1		-2 1		-3 1									
A	+1	2	0 10		-1 17		-2 9		-3 3							
A-	+2	2	+1 17		0 41		-1 71		-2 33		-3 12		-4 4			
B			+2 5		+1 42		0 114		-1 134		-2 49		-3 15		-4 **4**	*0.10*
C			+3 2		+2 16		+1 113		0 298		-1 274		-2 149		-3 **31**	*0.77*
D					+3 2		+2 30		+1 169		0 510		-1 739		-2 **241**	*5.98*
F			-				+3 1		+2 5		+1 59		0 1080		-1 **1236**	*30.67*
Total									**+3 5**		**+2 58**		**+1 402**		**0 2053**	*50.94*
%									*0.12*		*1.44*		*9.98*		*50.94*	-

Note. Subscript numbers = number of grade changes from the initial marking to the 1st remarking

Table 3
Comparison of the Grades obtained from the Initial Marking and 2nd Remarking

	Grades from 2nd remarking															
Grade	A+		A		A-		B		C		D		F		Total	
Initial marking	N	%	N	%	N	%	N	%	N	%	N	%	N	%		%
A+			1	*33.3*	1	*33.3*	1	*33.3*							3	*.1*
A	1	*2.4*	10	*24.4*	16	*39.0*	12	*29.3*	1	*2.4*	1	*2.4*			41	*1.0*
A-			14	*7.8*	49	*2.27*	67	*37.2*	30	*16.7*	15	*8.3*	5	*2.8*	180	*4.5*
B	1	*0.3*	3	*0.8*	45	*12.5*	113	*31.5*	121	*33.7*	63	*17.5*	13	*3.6*	359	*8.9*
C			2	*0.2*	25	*2.9*	136	*16.0*	292	*34.3*	272	*31.9*	125	*14.7*	852	*21.1*
D					6	*0.4*	49	*3.4*	237	*16.3*	511	*35.2*	647	*44.6*	1450	*36.0*
F			-				1	*0.1*	17	*1.5*	95	*8.3*	1032	*90.1*	1145	*28.4*
Total	2	*0.05*	30	*0.7*	142	*3.5*	379	*9.4*	698	*17.3*	957	*23.7*	1822	*45.2*	4030	*100*

Table 4
Change of Grades from Initial Marking to the 2nd Remarking

Grade Initial marking	Grades from 2nd remarking A+ N	A N	A- N	B N	C N	D N	F N	Total	%
A+	0	-1 1	-2 1	-3 1					
A	+1	0 10	-1 16	-2 12	-3 1	-4 1			
A-	1 +2	+1 14	0 49	-1 67	-2 30	-3 15	-4 5		
B	+3	+2 3	+1 45	0 113	-1 121	-2 63	-3 13	-4 **6**	*0.15*
C	1	+3 2	+2 25	+1 136	0 292	-1 272	-2 125	-3 **30**	*0.74*
D			+3 6	+2 49	+1 237	0 511	-1 647	-2 **231**	*5.73*
F		-		+3 1	+2 17	+1 95	0 1032	-1 **1124**	*27.89*
Total					+3 **10**	+2 **94**	+1 **528**	0 **2007**	*49.80*
%					*0.25*	*2.33*	*13.10*	*49.80*	-

Table *5*
Comparison of Grades obtained from the 1st and 2nd Remarking

Grade 1st re-marking	Grades from 2nd remarking A+ N	%	A N	%	A- N	%	B N	%	C N	%	D N	%	F N	%	Total N	%
A+					3	*75*	1	*25*							4	*0.1*
A			11	*31.4*	16	*45.7*	5	*14.3*	3	*8.6*					35	*0.9*
A-	1	*0.8*	8	*6.7*	39	*32.8*	50	*42.0*	16	*13.4*	5	*4.2*			119	*3.0*
B	1	*0.3*	7	*5.9*	62	*18.3*	129	*38.1*	97	*28.6*	34	*10.0*	9	*2.7*	339	*8.4*
C			4	*0.6*	18	*2.8*	134	*20.9*	246	*38.3*	187	*29.1*	53	*8.3*	642	*15.9*
D					3	*0.3*	49	*5.4*	239	*26.4*	382	*42.3*	231	*25.6*	904	*22.4*
F					1	*0.1*	11	*0.6*	97	*4.9*	349	*17.6*	1529	*77.0*	1987	*49.3*
Total	2	*0.0*	30	*0.7*	142	*3.5*	379	*9.4*	698	*17.3*	957	*23.7*	1822	*45.2*	4030	*100*

Table 6
Change of Grades from 1st to 2nd Remarking

1st re-marking	Grades from 2nd remarking A+ N	A N	A- N	B N	C N	D N	F N	Total N	%
A+			-2 3	-3 1					
A		0 11	-1 16	-2 5	-3 3				
A-	+2 1	+1 8	0 39	-1 50	-2 16	-3 5			
B	+3 1	+2 7	+1 62	0 129	-1 97	-2 34	-3 9		
C		+3 4	+2 18	+1 134	0 246	-1 187	-2 53	-3 **18**	*0.45*
D			+3 3	+2 49	+1 239	0 382	-1 231	-2 **111**	*2.75*
F		-	+4 1	+3 11	+2 97	+1 349	0 1529	-1 **581**	*14.42*
Total				+4 **1**	+3 **19**	+2 **172**	+1 **792**	0 **2336**	*57.97*
%				*0.02*	*0.47*	*4.27*	*19.65*	*57.97*	-

Table 7
Correlations between Marks and Grades for the Different Markings (N= 4030)

		Initial marking		1st remarking		2nd remarking	
		Marks	Grades	Marks	Grades	Marks	Grades
Initial marking	Pearson Correlation	1	1	.812	.744	.783	.713
1st remarking	Pearson Correlation	.812	.744	1	1	.829	.753
2nd remarking	Pearson Correlation	.783	.713	.829	.753	1	1

Table 8
Correlations of Marks from the Purposive Sample (N= 58)

		Initial marking	1st remarking	2nd remarking	Expert marks
Initial marking	Pearson Correlation	1	.662	.325	.484
1st remarking	Pearson Correlation	.662	1	.344	.562
2nd remarking	Pearson Correlation	.325	.344	1	.833
Expert marks	Pearson Correlation	.484	.562	.833	1

Interpretation of the Findings

From the 1st remarking (different between the initial grades and those obtained from the first remarking study), Table 1 shows a substantial change in grades awarded. In total, 1512 (37.5%) candidates obtained reduced grades and 465 candidates (11.5%) were upgraded. This means that a staggering 49% of candidates received a different grade on remarking. Also, the failure rate was increased from 28.4% to 49.3% illustrating the large degree of 'over-marking' in the actual examination to ensure candidates did not fail if their marks were close to the borderline mark.

On remarking the second time (the difference between the initial grades and those obtained on 2nd remarking), Table 3 shows that, in total, 1391 (34.5%) candidates obtained reduced grades. However, 632 (15.7%) candidates were upgraded meaning 50.2% candidates received a different grade on remarking. Also, the failure rate increased from 28.4% to 45.2%.

The difference between the two sets of remarking is also compared and this is given in Table 5. Table 5 indicates that from the 1st remarking to the 2nd remarking, 710 (17.6%) candidates obtained reduced grades. However, 984 (24.4%) candidates were upgraded. In total, 42% candidates received a different grade on remarking. Also, the failure rate decreased from 49.3% to 45.2%.

The Pearson product moment correlation between marks and grades obtained in the initial marking, 1st remarking and 2nd remarking are given in Table 7. These correlations are not particularly high, the highest being that between marks for the 1st and 2nd remarking (.829).

The correlations for a purposive sample of 58 answer scripts additionally marked by an expert for the examination board are given in Table 8. The data shows the highest

correlation with the sample from the 2nd remarking, illustrating that the raters in the 2nd remarking were closer to the intended marking direction.

Tables 2, 4, and 6 indicate actual grade changes from the initial to 1st remarking, initial to 2nd remarking and 1st to 2nd remarking respectively. The Tables also indicate the magnitude of the grade changes. Specific grade change magnitudes peak at -4 in Tables 2 and 4 and -3 in Table 6, supporting the notion that the 1st and 2nd remarking are closer than the grades given in the initial marking.

A paired comparison of the means, based on the grade obtained in the original marking, was undertaken to determine whether the difference in the marking, from the initial marking, the 1st remarking and the 2nd remarking, were significant. The results are shown in Table 9. The analysis illustrates that the change of grades on remarking between the initial grades awarded in the actual examination and the grades obtained on 1st remarking were significant ($p < 0.001$) for candidates in all grades, except for the few candidates who obtained a grade of A+ on the initial marking. To obtain this data, mean candidate grades on the 1st and 2nd remarking were compared with original grades in the initial examination on a grade by grade basis.

The Rasch analysis Figures (1-5) show the range of severity among the raters. Figure 1 derives from scripts re-marked randomly by 11 raters. It shows that candidates (indicated by x's and with examination score totals remarked by the 11 raters) are of relatively low ability, lower than the arbitrary 0 on the logit scale. In the remarking, raters 5 and 7 were the most severe and rater 9, the most lenient. The spread of severity of inter-rater marking is quite large and thus the unity of inter-rater marking relatively poor. Figure 2 replicates Figure 1, but applies to the inserting of marks of the same scripts in a 2nd remarking, with the 11 raters marking similar numbers of scripts, but no longer marking, for the most part, the same scripts as in Figure 1. The range in severity of the raters is shown to be much less, although rater 5 is still the most severe and rater 9, one of the most lenient. Figure 3 derives from the same random set of scripts, but marked by the 11 raters in a 2nd remarking after receiving guidance training. Unfortunately the diversity in severity of marker is now much greater with rater 7 becoming much more severe and rater 1, much more lenient. Figure 4 replicates Figure 3, but relates to the insertion of marks on the same scripts from the 1st remarking, but by the 11 raters now marking, for the most part, different scripts. Here the range of rater severity is shown to be much less and indicates that the removing the influence of the training effect leads to more interrater agreement. Figure 5 shows the 3 marking outcomes put together as undertaken by the 11 raters, showing overall the range of severity is reduced from rater allocations on the first marking and that raters are moving towards some commonality. Unfortunately the differences are not diminished following the 2-day training programme; in fact if anything it has increased the differences among the raters! This finding is in agreement with that found by Stahl and Lunz (1991) and Weigle (1998). It seems that such training, not linked to a purposive activity in the eyes of the raters, is not particularly useful. The separation reliability is very high >.98, indicating that the separation of the rates into different levels is very reliable. And with significant values for chi-squared, the ordering of candidates by raters is not constant with the estimated ability measure of candidates.

The examination also clearly shows that the questions, as a whole, were not answered well and presumably these questions were far different from those expected by the candidates. This clearly suggests a construct validity issue. If the

questions have really been developed at the appropriate level, then the teaching in schools is very much out of step. Further, recognising that the question setters are teachers from another district of the country, there is the suggestion that expectations differ from examination board to examination board within the country. And if this is the case, then examination boards would seem to have little worth.

Table 9
Paired Samples Test of Significant Differences between Means Marks based on Initial Grades

Comparisons	Paired Differences					*t*	*df*	*P*
	M	*SD*	*SE*	95% Confidence Interval of the Difference Lower	Upper			
1st remarking mean marks compared with initial A+ grade marks	8.00	4.36	2.52	-2.83	18.83	3.18	2	.086
2nd remarking mean marks compared with initial A+ grade marks	7.00	3.61	2.08	-1.96	15.96	3.36	2	.078
1st remarking mean marks compared with initial A grade marks	3.93	4.94	0.77	2.37	5.49	5.09	40	.000
2nd remarking mean marks compared with initial A grade marks	4.49	4.75	0.74	2.99	5.99	6.05	40	.000
1st remarking mean marks compared with initial A- grade marks	4.01	5.56	0.41	3.19	4.82	9.67	179	.000
2nd remarking mean marks compared with initial A- grade marks	4.49	5.89	0.44	3.63	5.36	10.24	179	.000
1st remarking mean marks compared with initial B grade marks	2.73	4.82	0.25	2.23	3.23	10.73	358	.000
2nd remarking mean marks compared with initial B grade marks	2.77	4.90	0.26	2.26	3.28	10.71	358	.000
1st remarking mean marks compared with initial C grade marks	2.39	5.05	0.17	2.05	2.73	13.83	851	.000
2nd remarking mean marks compared with initial C grade marks	1.85	5.17	0.18	1.50	2.19	10.43	851	.000
1st remarking mean marks compared with initial D grade marks	2.29	4.92	0.13	2.04	2.55	17.77	1449	.000
2nd remarking mean marks compared with initial D grade marks	1.45	5.33	0.14	1.18	1.73	10.40	1449	.000
1st remarking mean marks compared with initial F grade marks	-0.32	3.88	0.11	-0.55	-0.10	-2.87	1144	.004
2nd remarking mean marks compared with initial F grade marks	-1.06	4.35	0.13	-1.32	-0.81	-8.28	1144	.000

Figure 1
Rater effect on 1st Remarking
Map of Latent Distributions and Response Model Parameter Estimates

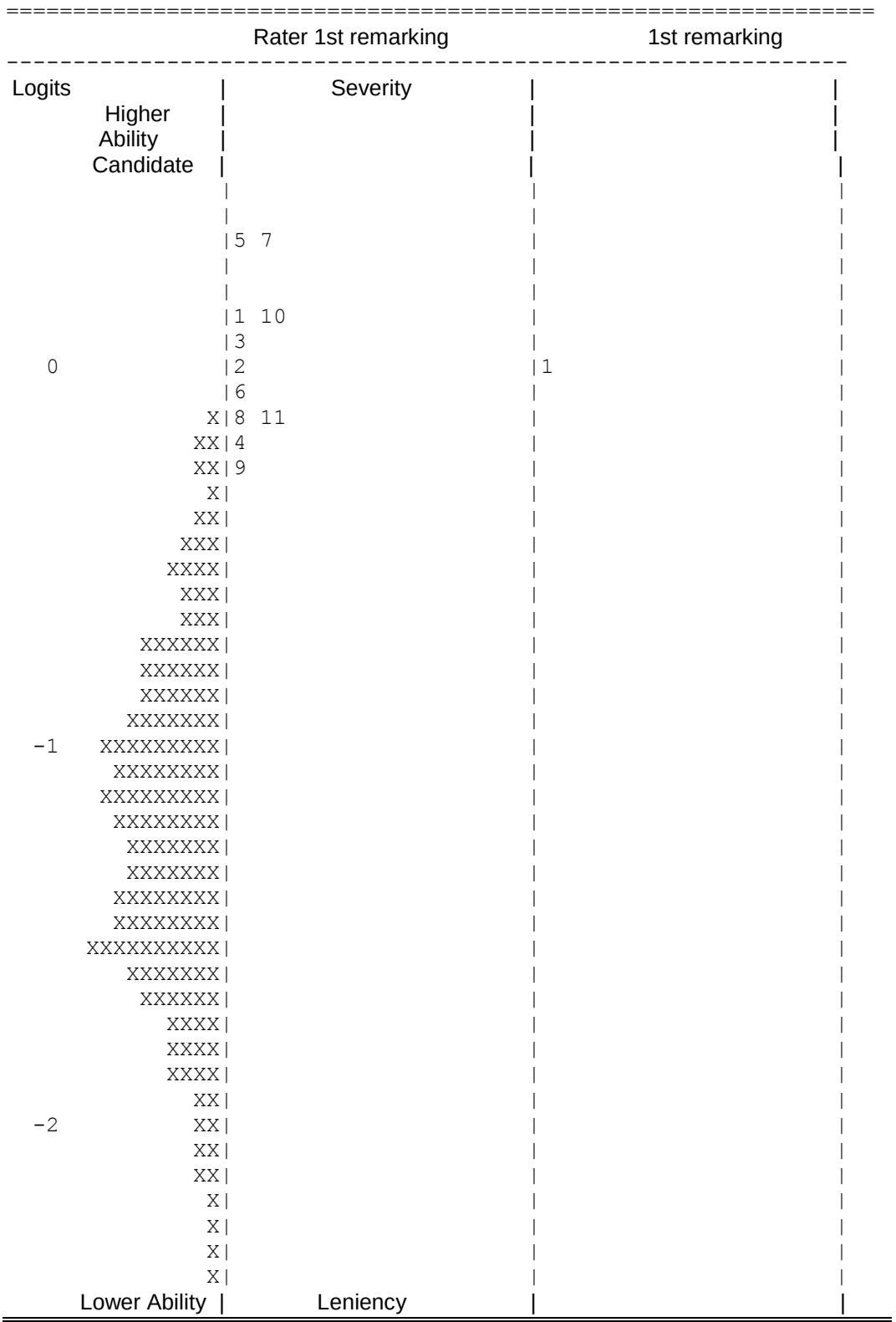
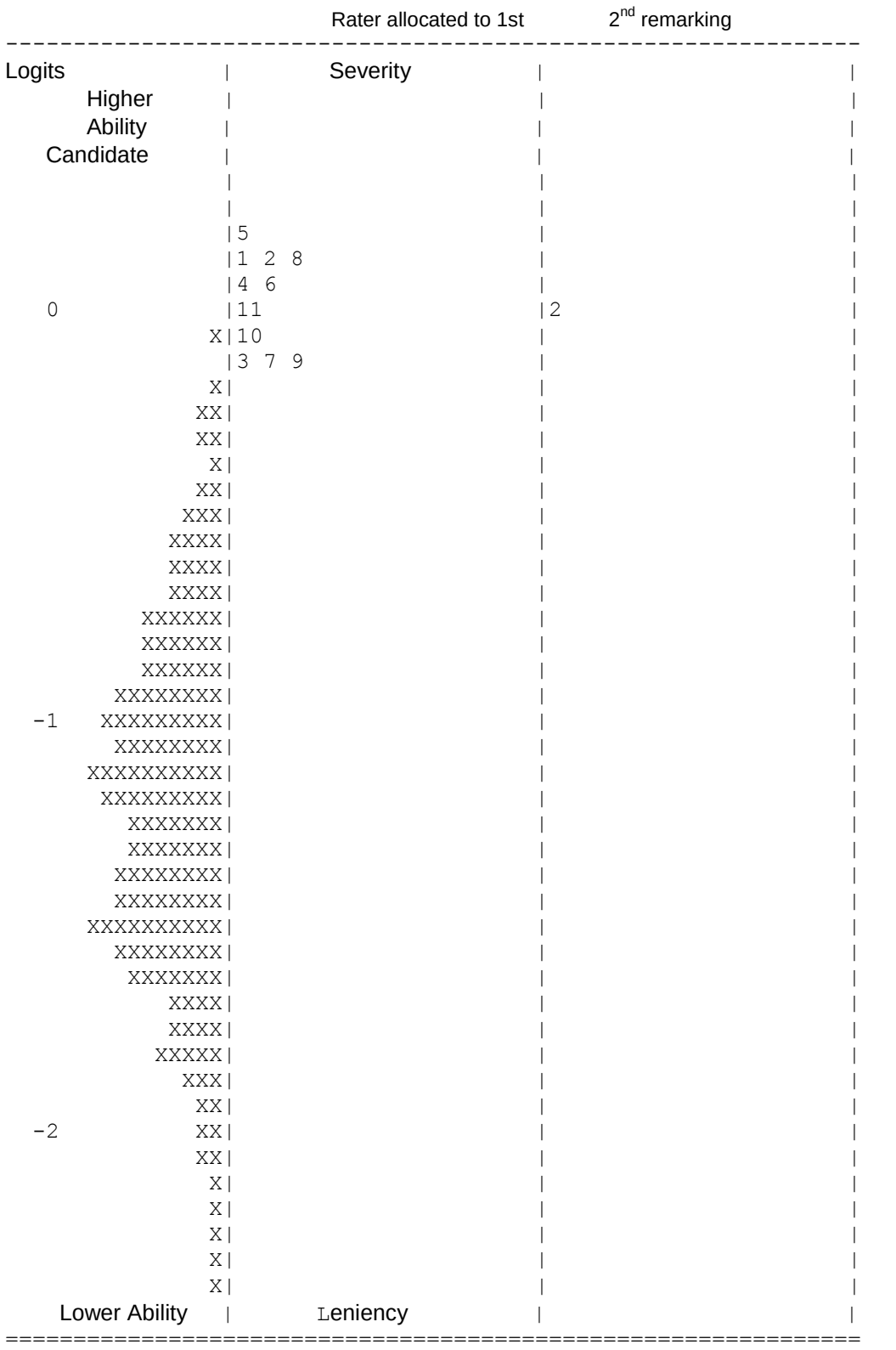


Figure 2
Rater effect on 2nd Remarking
Map of Latent Distributions and Response Model Parameter Estimates



Each 'x' represents 23.9 candidates

Figure 3
Rater effect on 2nd Remarking
Map of Latent Distributions and Response Model Parameter Estimates

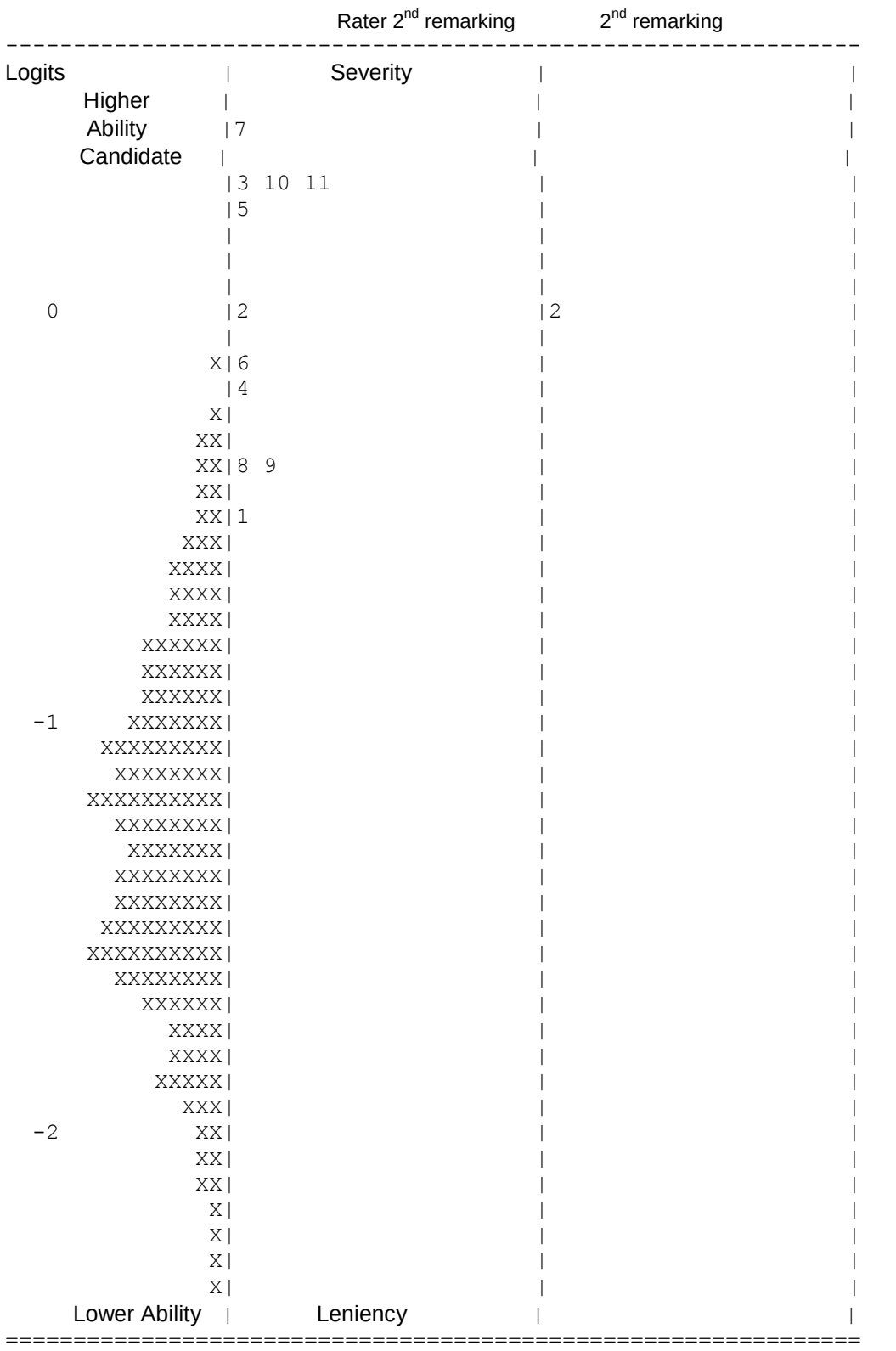


Figure 4
Rater Effect on 1st Remarking
Map of Latent Distributions and Response Model Parameter Estimates

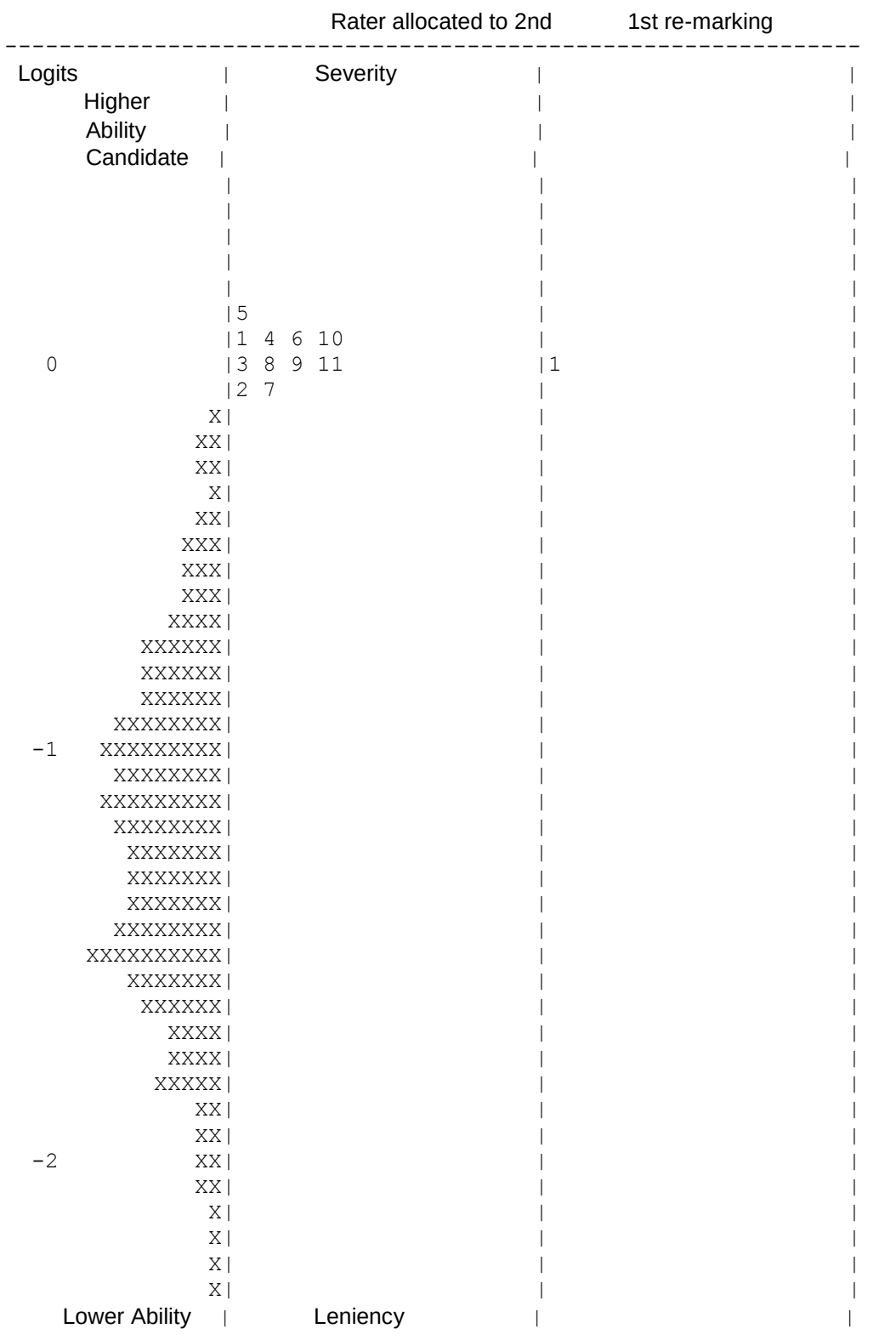
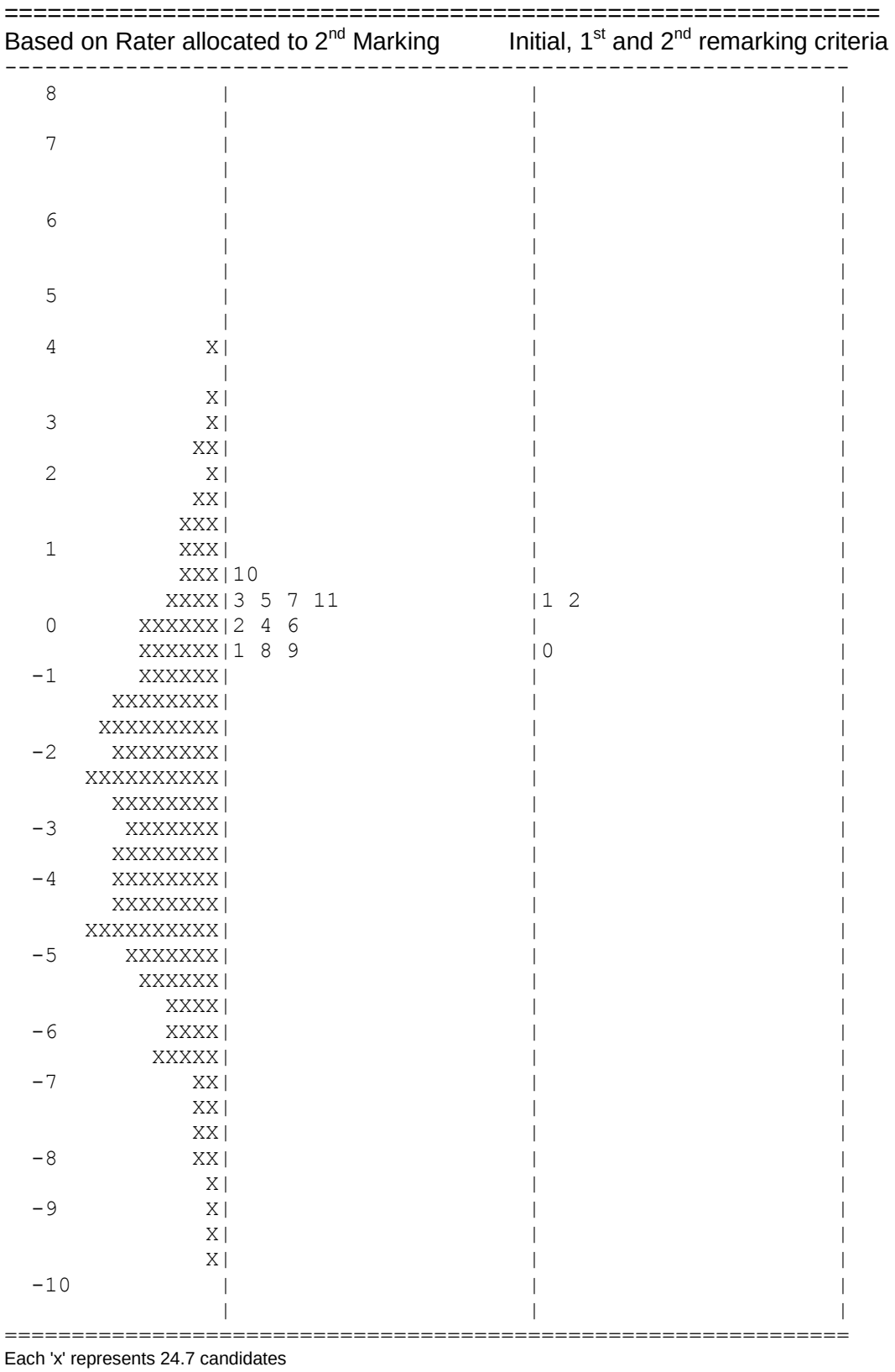


Figure 5
Rater Effect on 3 Levels of Marking
Map of Latent Distributions and Response Model Parameter Estimates



Each 'x' represents 24.7 candidates

Discussion

The remarking gave significantly different total marks and different grades at all levels except for those few candidate who obtained a grade of A+. It is thus clear that the hypothesis put forward stipulating there is no significant change in grades on remarking is found to be incorrect. There are change of grades for a significant number of candidates (except A+ grades). This of course is very undesirably, as it underminds the validity of the whole examination. The legitimisation of setting up and running expensive examination boards, to run high stakes examinations, is under threat. It is importanrt that steps be put in place to redress this concern.

Clearly the marking of these answer scripts is very unstable and there is little reliability in the marks obtained. It is staggering that remarking can change the grade awarded by as much as 4 grade points in some cases (see Tables 1 and 2 where (i) the grades for 4/5 candidates were changed on remarking from A- to F in each case and (ii) 1 candidate, on remarking, changed grade from F to B). Furthermore this highlights the chance of personal error/bias in marking and the minimal student-teacher relationship regarding the question content. It is thus clear that the study points to the need for a more reliable marking system (if candidate total marks as well as grades, are to be taken as standard and there is a need to minimise personal bias /error and encourage better teaching-learning situations in institutions). It is suggested that detailed marking guidelines (and even full answers where questions tend towards a more structured format) are produced by question setters, not by head examiners. In this way, the construct validity of the questions is enhanced, because the marks are more likely related to the intentions of the questions. Such detailed mark schemes will also assist question paper moderators better understand the questions set and again aid the validity of the question paper.

The need to develop a more reliable marking system suggests that better guidance to raters is required, illustrating how to mark candidate answer scripts more effectively. It is suggested that more reliable marking can be obtained by:

(a) presenting each examiner with a detailed marking scheme with actual answers for marking the script so that it becomes easier to award marks to common criteria;

(b) setting questions that better lend themselves to more objective and hence more specific marking schemes. This means, for subjects where creativity and presentation are less important than the expression of conceptual ideas, moving away from essay-type questions with its subjective marking procedures and towards questions which have a set structure and which can be used to award marks in a more objective manner. Such questions can be structured questions which are broken down into sub-parts and each part is marked separately based on a set of specific criteria (Verna, Chhatal, & Singh, 1997);

(c) requiring all examiners to mark initially the same set of answer scripts (sample marking using 6-9 answers scripts across the mark range) and then discussing the marks awarded to determine a common set of detailed marking guidelines for all raters.

One approach suggested earlier is to consider double marking by averaging marks from the initial and first remarking. Table 8 compares the grades awarded by

the averaging with those from a third marking (the 2nd remarking). As the 2nd remarking was carried out after a 2 day in-sevice course, these marks can be expected to be more authentic. This is supported by a correlation of marks (Table 9) from a purposive sample of 58 scripts in which the the grades awarded by all 3 forms of marking were compared with the marks awarded by an 'expert' familiar with the questions and the answers expected. The finding suggest, however, that double marking is not likely to be as effective. Add to this the time and cost factor (Newton, 1996; Verma, Chhatwal, & Singh, 1997) and clearly double marking is not really as appropriate as taking steps to marking the questions more objectively and mandating a detailed marking scheme.

Conclusion

The examination questions are poorly marked, with far too great a leniency for candidates with lower marks. The marking improved on the remarking, but the guidance given before raters attempted the second remarking did not show substantial gains. A comparison of the rater leniency and harshness changed little. The inter-rater reliabliity of the examination under scrutiny was found to be low. This is alarming given that the examination is high stakes and the grades awarded to candidates means much in determining their future. Clearly the examiners were given insufficient guidance as shown by the greater correlation with an expert rater when marking after a 2 day in-service course. But the lack of agreement in the marking of all questions suggests that examiners do need a detailed answer script, besides practice in its use by marking of sample scripts.

Noting that most candidates obtained grades in the middle of the range and that even in this range there were substantial discrepancies between grades awarded on remarking, inter-rater consistency must be a great cause for concern. Examination boards clearly need to give much more attention to using raters who are prepared for the task and who recognise the importance of reliable marking. The need for examination boards is being undermined by the lack of inter-rater consistency.

This was an initial study into reliability of the marking and shows that all 3 hypotheses need to be rejected, although there is some evidence that limited in-service training did lead to more appropriate marking. Further studies can be undertaken such as the effects of motivation and rewards given to the raters and the careful scrutinising of marks after being recorded on the candidate answer scripts. These are all matters, suggested by raters, to be of concern.

References

Altman, D. G., & Bland, J. M. (1983). Measurement in medicine: The analysis of method comparison studies. *The Statistician*, *32*(3), 307-317.

Baird, J-A., Greatorex, J., & Bell, J. P. (2004). What makes marking reliable? Experiments with UK Examinations. *Assessment in Education*, *11*(3), 331-348.

Bédard, M., Martin, N., Kruger, P., & Brazil, B. (2000). Assessing reproducibility of data obtained with instruments based on continuous measurements. *Experimental Aging Research*, *26*, 353-365.

Berkowitz, D., Wolkowitz, B., Firch, R., & Kopriva, R. (2000). *The use of tests as part of high-stakes decisions for students: A resource guide for educators and policy-makes*. Washington DC: US Dept. of Education.

Brennan, R. L., & Prediger, D. J. (1981). Coefficient kappa: Some uses, misuses, and alternatives. *Educational and Psychological Measurement*, *20*, 37-46.

Brooks, V. (2004). Double marking revisited. *British Journal of Educational Studies*, *52*(1), 29-46.

Campbell, A. (2005). Application of ICT and rubrics to the assessment process where progessional judgement is involved: the features of an e-marking tool. *Assessment and Evaluation in Higher Education*, *30*(5), 529-537.

Chase, C. I. (1986). Essay test scoring: Interaction of relevant variables. *Journal of Educational Measurement, 23*, 33-42.

Crossley, J., Davies, H., Humphris, G., & Jolly, B. (2002). Generalisability: a key to unlock professional assessment, *Medical Education*, *36*, 972-978.

Dahl, T .I. (2006). When precedence sets a bad example for reform: conceptions and reliability of a questionable high stakes assessment practice in Norwegian universities. *Assessment in Education*, *13*(1), 5-27.

Dimitrov, D. M. (2002). Reliability: Arguments for multiple perspectives and potential problems with generalization across studies. *Educational Psychological Measurement, 62*(5), 783-801.

Eckes, T. (2005). Examining rater effects in TestDaF writing and speaking performance assessment: A many facet Rasch analysis. *Language Assessment Quarterly*, *2*(3), 197-221.

Elder, C., Knoch, U., Barkhuizen, G., & von Randow, J. (2005). Individual feedback to enhance rater training. Does it work? *Language Assessment Quarterly*, *2*(3), 175-196.

Engelhard, G. Jr. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement*, *31*(2), 93-112.

Lunz, M., Wright, B., & Linacre, J. (1990). Measuring the impact of judge severity on examination scores. *Applied Measurement in Education*, *3*(4), 331-345.

Marsh, H. W., & Ireland, R. (1987). The assessment of writing effectiveness: a multidimensional perspective. *Australian Journal of Psychology*, *39*, 353-367.

Moon, T. R., & Hughes, K. R. (2002). Training and scoring issues involved in large-scale writing assessments. *Educational Measurement: Issues and Practice*, *21*(2), 15-19.

Munro, N., Denney, M. L., Nughani, A., Foulkes, J., Wilson, A., & Tate, P. (2003). Ensuring reliability in UK written tests of general practice: The MRCGP examination 1998-2003. *Medical Teacher*, *27*(1), 2005, 37–45.

Newton, P. E. (1996). The reliability of marking of general certificate of secondary education scripts: Mathematics and English. *British Educational Research Journal*, *22*(4), 405-420.

Nuttall. D. L., & Willmott. A.S. (1972). *British examinations: Techniques of analysis P12*. Slough: NFER Publishing.

Van der Vleuten, C. P. M. (1996). The assessment of professional competence: Developments, resources and practical implications. *Advances in Health Sciences Education, 1*, 41-67.

Verma, M., Chhatwal, J., & Singh T. (1997). Reliability of essay type questions--Effect of structuring, *Assessment in Education*, 4(2), 265-71.

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing, 15,* 263–287.

Wu, M. L., Adams, R. J. & Wilson, M. R. (1998). ACER ConQuest. Melbourne, ACER.

Yen, W. M. & Candell, G. L. (1991). Increasing score reliability with item-pattern scoring: An empirical study in five score metrics. *Applied Measurement in Education*, *4*(3), 209-228.

Jack Holbrook is a visiting professor in science education at the University of Tartu, Estonia and an independent consultant, specialising in assessment, curriculum and teacher education at the primary and seocndary level. He has a PhD in Chemistry from the University of London and was trained as a secondary school teacher before spending most of his career in teacher education – in the UK, Africa and Asia. His main area of research is in the philosophy of science education and relating this to the teaching of science subjects. He is the current President of the International Council of Associations for Science Education (ICASE).