The *Zimbabwe Bulletin of Teacher Education* is published three times a year by the University of Zimbabwe, Department of Teacher Education, Faculty of Education.

ISSN No–1022–3800

Department of Teacher Education
Faculty of Education
University of Zimbabwe
P O Box MP 167
Mount Pleasant
Harare
Zimbabwe

The Zimbabwe Bulletin of Teacher Education

Volume 4 No 1, March 1994

RELIABILITY IN THE MARKING OF DISTANCE EDUCATION EXAMINATIONS

Fred Zindi
Department of Educational Foundations
University of Zimbabwe

ABSTRACT:

Forty examiners met at the ZIPAM Centre in Zimbabwe to mark scripts produced by Bachelor of Education (Administration) students who had taken distance education courses in Research Methods, Schools Management, Leadership and Supervision, Educational Planning and Policy Studies.

Initial training of the examiners was given through lectures, seminars and the marking of "dummy" scripts in order to address the question of marker reliability and to standardize the marking procedure.

After lengthy discussions between markers and moderators of the "dummy scripts" an attempt to iron out differences between those markers who gave a wide range of scores for the same scripts was made.

Using revised and more elaborate marking schemes the examiners, who were continuously monitored after marking every four scripts, were then asked to mark the 'live' scripts.

The results indicated that there was no complete reliability in the marking. There was, however, a slight increase in the overall reliability of marking noticed between the initial stage of 'dummy' marking and the final stage of 'live' marking which occurred after the brainstorming session.

Background

As far back as 1935, Hartog and Rhodes conducted a study to investigate the fairness and consistency of marking essay-type examinations. They selected 15 history scripts which had been given the same

middle-of-the-road marks by a panel of 15 markers, all of whom were experienced examiners. A year later, the same essays were marked again by the same examiners inorder to provide an indication of test-retest reliability of the marking. On nearly 50 per cent of the scripts the examiners had changed the marks they had awarded a year earlier, with some students who had been passed the first time, failing when the same scripts were marked by the same markers on the second occasion. In one particular case, an examiner awarded the same script marks which differed by 30 points.

Since the Hartog and Rhodes study, several other studies have been conducted over the years. Nisbet (1971) designed a framework for student achievement which consisted of three principal approaches:

(i) a content framework where the student's score is given as a percentage of a defined sample of knowledge and skills that have been learned;

(ii) an absolute framework in which the student's score indicates the levels of achievement attained as specified in the course objectives; and (iii) a normative framework where the student's performance is measured against the attainment of others.

Nisbet recommended the understanding of the statistics of marking using marking guides in order to avoid confusing the meanings of the above frameworks. This way, according to Nisbert, instinct may then legitimately be transformed into the qualitative professional judgement necessary for the crucial responsibilities teachers have when assessing students.

There is no doubt that marking procedures are often taken for granted, with a minimum of rational analysis, especially when marking involves essay-type questions. It is easy to be subjective, yet the outcome of such marking will often determine the student's future such as dismissal from an institution or being asked to withdraw from courses. It is such actions of subjectivity that account for much of the irrationality of marking

practices. For instance a marker who may wish to fortify his ego with the conviction that "standards must be maintained" can take satisfaction in the high rate of failure and the infrequency of grade As in his classes. Another marker may decide to give As only to his students with the hope that his colleagues may see this as evidence of good teaching on his part.

There are several other factors that can influence the marking of examination scripts. The marker who knows the identity of the students whose scripts he is marking could be influenced by factors such as sex, ethnic background, previous test scores or the physical characteristics of the students. Alternatively, subjectivity could arise through the pygmalion effect or teacher-expectancy effect.

How can examiners therefore ensure that their marking is reliable? In order to ensure that the mark awarded is fair one, a number of checks are often used. These include the double marking of scripts (by two different examiners) and the use of external examiners at the final stage of the marking in order to ensure that standards are the same. However, even when such checks are in place, marker unreliability often persists.

Thorndike and Hagen (1977), define reliability as "the accuracy and precision of a measurement procedure." A completely reliable marking system therefore, is one from which a student whose scores at a particular level are always the same. However, this is not always practical as student performance fluctuates from time to time and may also vary from question to question. Markers, as discussed above, may give different scores when evaluating the same answers, or the same marker becomes inconsistent when evaluating the same scripts on different occasions, thus creating unreliability.

Apart from double marking of scripts, national examination boards often recommend the use of elaborate and detailed marking schemes as well as discussions among markers on the criteria to used for awarding marks. Murphy (1982) gave similar examiners scripts that had been marked by other examiners and found that after the scripts had been re-marked,

near-perfect inter-marker correlations of around 0,9 were obtained. It would therefore appear that such measures which include meetings and address the comparability of standards have had some degree of success.

The present study investigated the reliability of the marks awarded by Distance education examiners to examination scripts produced by Bachelor of Education (Administration) students.

Method

Forty markers drawn form all distance education centres throughout Zimbabwe met at the ZIPAM Centre to participate in the marking of examination scripts. All markers were experienced examiners, but as part of an exercise on marking standards the first two days were spent on training the markers on ways of standardizing the marking as well as maintaining reliability. Lectures, discussions and the marking of dummy scripts were conducted during the training period. The 40 markers were then placed in five groups, each with eight members. Each group was given four dummy scripts in their specialist subject areas. For instance, group A was given four scripts in Research Methods, group B in Schools Management, group C in Leadership and Supervision, Group D in Educational Planning, while group E dealt with Policy Studies. Each script was then marked eight times by the eight markers in each group.

Each script had three essays written in three hours from a choice of seven questions. The 'dummy' scripts which were marked 'blind' were photocopies of essays written under real examination conditions but with students' names and identity numbers removed. The markers were asked to mark on the normal percentage scale and criteria used by the University of Zimbabwe (0 - 49 = Fail; 50 - 59 = Third; 60 -69 = Lower Second; 70 - 79 = Upper Second; 80 - 100 = First).

This was followed by a comparison of mean scores between markers. Table 1. below shows the means, ranges and standard deviations of marks awarded by examiners in each group on the dummy scripts.

Table 1: Means, Ranges and Standard Deviations of Scores Awarded by 5 Groups of Markers on 4 Dummy Scripts.

GROUP A Research Methods				GROUP B Managing Schools			GROUP C Leadership & Supervision			GROUP D Planning & Development			GROUP E POLICY STUDIES		
St	Mn	Rg	SD	Mean	Range	S.D.	Mean	Range	S.D.	Mean	Range	S.D.	Mean	Range	S.D
1	61.8	14	4.2	58.3	11	3.5	63.5	9	2.1	52.5	13	6.7	70.1	11	3.1
2	59.8	27	7.7	70.2	18	6.3	57.8	23	7.3	62.7	23	5.8	63.5	17	4.8
3	69.8	16	5.6	59.8	21	4.5	52.5	16	4.4	62.2	15	5.5	51.2	19	9.1
4	73.5	19	6.3		36	6.7	55.9	31	4.7	66.6	12	6.6	59.7	18	7.6

KEY

St - Student

Mn - Mean

Rg - Range

SD - Standard Deviations

The marks of each group were also subjected to a two-way analysis of variance in which the factors were markers (8) and scripts (4). The mean squares from this analysis allow an estimate to be made of a number of components of variance. The standard error of measurement was 7.2 in group A, 7.1 in group B, 6.3 in group c, 5.7 in group D and 6.6 in group E.

Scores were also analysed to see if they could provide any insight into the reasons for the differences between markers. One possibility was that the markers agreed quite well in how they rated a particular script on different dimensions, but disagreed as to how these dimensions should be weighted to arrive at a final mark.

An analysis of variance was applied to the ratings between markers in order to look at interaction between scripts and attributes. This interaction was not significant ($F = 0,93$, $p = 0,08$) suggesting that markers did not have a common view of where the strengths and weaknesses of each script were. The markers who rated a particular script highly were also found to rate the rest of the scripts they marked quite highly and vice-versa for those who gave low marks.

After discussion and brainstorming the marking scheme was improved before the marking of 'live' scripts began. Once again, although there was a marked improvement in the overall reliability of marking (the average standard error of measurement observed was 4.45 i.e. less than before) there was still considerable variation between markers. The improvement in overall reliability was attributed to the elaborate procedures followed to ensure reliability and comparability of standards. These included the training given in assessment procedures, detailed marking schemes and discussions which followed after the marking of 'dummy' scripts.

Table 2 shows that the distribution of scores after the marking of 90 live scripts in the 5 examinations seem to follow a normal distribution and an increase in reliability from the dummy script marking.

Table 2: Scores Distribution in 5 Exams

Score	EXAM 1	EXAM 2	EXAM 3	EXAM 4	EXAM 5	TOTAL
80-100	1	2	1	1	1	6
70-79	2	4	4	5	3	18
60-69	4	3	3	6	4	20
50-59	8	7	9	10	8	42
0-49	1	0	1	0	2	4
N	16	16	18	18	18	90

Discussion

The findings of this study compare very well with those of Dennis and Newstead (1994) who also found that markers disagree among themselves as to what individual pieces of work are worth. Despite the fact that examiners are charged with the responsibility of ensuring that equivalent standards are adopted when marking scripts based on the same questions, the variation in opinions is particularly disturbing since these opinions carry great weight in determining the final classification of students.

Although it was found that there was a certain level of disagreement among markers, this does not mean that students who write essay type examinations are constantly being incorrectly classified. It is usually those students who are close to the borderline who are likely to be misclassified and these are typically the students who are considered at length by external examiners. It is well-documented that the probability of an incorrect overall classification through marker unreliability is actually quite small (Dennis and Newstead, 1994). As a result despite these small discrepancies, the current classification system continues to be used extensively.

With continued monitoring of markers especially the "too harsh" and the "too lenient" markers, examiner bias and unreliability would decrease tremendously if, in addition, the steps suggested above in the method section are also followed. The more separate assessments and the more markers that contribute to the final mark, the more reliable that final mark, will be. There are obviously some limits to the amount of marking that markers can reasonably be asked to do. The ideal thing would be to ask different markers to mark each script as many times as possible, discussing differences and then agreeing on a final mark, but in practice due to the shortage of manpower and the insufficient time given to carry out such an exercise, this is not always possible. Besides, this would be a very expensive exercise. What often happens, however, is a remark of extreme cases, such as those getting very high marks and those who fail.

In addition, all the evidence obtained in the ZIPAM exercise seemed to suggest that there was greater reliability when marking schemes were adopted although some of the more senior markers initially protested that marking schemes were too prescriptive and did not allow for those students who answered in unusual and creative ways.

Such cases are however, rare, and marking schemes could be used more flexibly and still continue to provide reliability in marking.

Recommendations

There is clearly room for improvement in our marking systems especially when examination questions are non-objective. Steps such as the training of markers, double marking of scripts and discussions should therefore be taken to ensure reliability.

During the ZIPAM exercise, there was a possibility that in some places, different markers could have interpreted the demands of the examination questions differently and thus causing the differences in their decisions as to what each script was worth. Future standardization exercises should

therefore begin with an analysis of the examination items since it is a known fact that item bias has an influence on both performance and the end result (Zindi; 1994).

There is also a need for markers, moderators and external examiners to read the same distance education modules which are supplied to the students (as opposed to the use of own already-existing knowledge base) before marking scripts in order to have a better understanding of how students arrive at their answers.

References

Dennis, I and Newstead, S. (1994) Examiners Examined. *The Psychologist,* 7 No. 5, p 216-219.

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

Murphy, R.J.L. (1982) A Further Report into the Reliability of Marking of GCE Examinations. *British Journal of Educational Psychology*, 52, p 58-68.

Nisbet J.D. (1971) Framework for Student Achievement. A paper submitted at the *British Educational Psychology Symposium.* London.

Thorndike, R.L. and Hagen, E.P. (1977) *Measurement and Evaluation in Psychology and Education* 4th Edition; New York: John Wiley and Sons.

Zindi, F. (1994) Differences in Psychometric Performance. *The Psychologist 9* No. 5, p 214-217.