MBOKOCHENA EDMORE, MUSHORIWA D. TARUVINGA & MUDZENGERERE F. HAMILTON

Prediction Levels of Teacher-Made Tests: A Case of Ordinary Level Geography in Highfield, Zimbabwe

ABSTRACT: *The study sought to establish prediction levels of teacher-made tests in final ordinary level examinations as a way of assessing the predictive validity of teacher-made tests. Specifically, the investigation focused on the extent to which teacher-made tests in Geography predict pupil performance in final examinations. Data was collected from six Geography teachers and mark records of sixty pupils selected from two secondary schools using stratified random sampling. This survey research design was used in this study for having a true representative sample. Positive and significant relationship between teacher made test results and final examination results were found in all cases. Significant differences in performance were also observed among the four streamed classes. Among the recommendations was the need to ensure that teachers are well-grounded in measurement theory in order to help them not only set valid tests, but also know how to reduced "error of measurement" when scoring tests.*
KEY WORD: *Final examination, prediction levels, secondary schools, teacher–made tests, geography, scores, and tests.*

IKHTISAR: *Artikel ini berjudul "Tingkatan Prediksi Tes Buatan Guru: Sebuah Kasus Tingkatan Biasa pada Bidang Geografi di Wilayah Highfield, Zimbabwe". Penelitian ini berusaha untuk menetapkan tingkatan prediksi tes buatan guru dalam ujian akhir tingkatan biasa sebagai cara untuk menilai validitas prediktif terhadap tes buatan guru. Secara khusus, penyelidikan difokuskan pada sejauh mana tes buatan guru dalam bidang Geografi dapat memprediksi kinerja murid dalam ujian akhir. Data dikumpulkan dari enam guru Geografi dan catatan buatan enam puluh murid terpilih dari dua sekolah menengah dengan menggunakan sampel acak bertingkat. Desain penelitian survei yang digunakan dalam penelitian ini dimaksudkan untuk memiliki sampel yang benar-benar representatif. Hubungan positif dan signifikan antara hasil tes buatan guru dan hasil ujian akhir ditemukan dalam semua kasus. Perbedaan yang signifikan dalam kinerja juga diamati antara empat kelas secara berurutan. Diantara rekomendasinya adalah perlu untuk memastikan bahwa guru cukup beralasan untuk melaksanakan teori pengukuran dalam rangka membantu mereka tidak hanya menata tes yang valid, tetapi juga mengetahui bagaimana mengurangi "kesalahan pengukuran" ketika menskor tes.*
KATA KUNCI: *Ujian akhir, tingkatan prediksi, sekolah menengah, tes buatan guru, pelajaran geografi, skor, dan tes.*

INTRODUCTION

The current study examined the extent to which teacher-made tests in Geography predict pupils' performance in the final examination at "O" level. Teacher-made tests are regularly administered by teachers, not only for checking pupils' progress and achievement but also as a means of preparing for final examinations (Mehrens & Lehmann, 1978). Many parents, teachers, pupils, and other stakeholders in education place a lot of faith in the potency and usefulness of teacher-made tests. These tests are viewed as providing focus, priority, direction, and pace for pupils' learning. Given this importance attached to teacher-made tests, to the present writers, surely there is a case for a better understanding of the real contribution of teacher-made tests to pupils' performance in the final examinations.

While large-scale studies on the validity of

Mbokochena Edmore, M.A. is a Lecturer at the Department of Sociology WUA (Women's University in Africa), Harare, Zimbabwe; **Mushoriwa D. Taruvinga, Ph.D.** is a Lecturer at the Department of Educational Foundations and Management UNISWA (University of Swaziland), Swaziland; and **Mudzengerere F. Hamilton, M.Sc.** is Research Fellow at the Faculty of the Built Environment NUST (National University of Science and Technology), Bulawayo, Zimbabwe. Corresponding author is: eddiembok@gmail.com

teacher-made tests have been conducted elsewhere, to the writers' knowledge, there is a facility of such research in Zimbabwe. Many researchers in Zimbabwe have tended to assume that what has been found in other countries, also applies to Zimbabwe, yet circumstances, environments, and education systems are different. It is perhaps for this reason that N. Mwatengahama (1999) says that there has not been much effort on the part of professionals in this country (Zimbabwe) to check on the validity and accuracy of teacher judgments of students' academic performance.

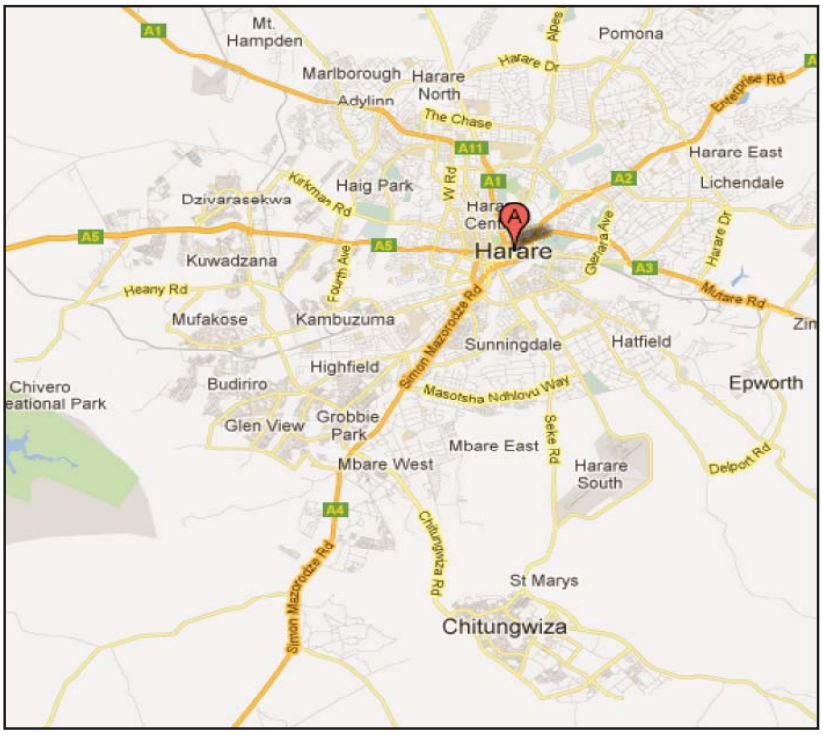

Figure 1:
The Map of Harare Showing Highfield, High Density Suburb
(Source: www.google/maps.com, 21/4/2013)

The current study, therefore, takes lofty significance considering the fact that little attention has been given to the issue of the accuracy and relevance of teacher-made tests in previous researches in this country. A study of this nature may generate research data that truly reflect and tell the Zimbabwean story regarding the role of testing in guiding teaching and learning.

SCOPE OF THE STUDY AND STUDY AREA

The research aims at analysing how in-class tests can be used to predict the final grade marks for students. It also seeks to address how consistent was the performance of students in Geography during the school examinations and final examinations. The research helps to improve on the board of knowledge and proving recommendations on how teachers can improve on pupils' performance.

The research was conducted in Highfield high density area of Harare in Zimbabwe. Highfield lies in the southern high density suburbs of Harare and has a high population density. Being one of the oldest locations in the capital city, the area is characterised by students with different academic abilities. Primary schools in the area feed the secondary schools with students. The diagram below shows the location of Highfield suburb in Harare, Zimbabwe.

RESEARCH METHOD

The survey research design was used, because it can be used to collect a great deal of information from a large population in a short period. Furthermore, survey manipulation allow the gathering of data at a particular point in time with the intention of describing the nature of existing conditions, or identifying standards against which existing conditions can be compared, or determining the relationship between specific events or phenomena (Kerlinger, 1986). It was the intention of the current study to determine the relationship between teacher-made test scores and final examination scores with a view to improving teacher-made tests. The survey design was thus facilitative in this respect.

The study involved two secondary schools in Highfield suburb of Harare. The total number of pupils who wrote both the mid-year test

and the final Zimbabwe School Examination Council (ZIMSEC) examination in Geography in November 2004 was 366. Stratified random sampling was used in this study since the two schools streamed their pupils according to ability.

In each school, there were four Form IV classes, with Class A as the best, followed by classes B, C, and D. Two classes were selected from each school through stratified random sampling. From School Z, classes 4A and 4D were selected; while from School Y, classes 4B and 4C were selected. This ensured that mark profiles of pupils of different abilities were represented in the sample. The researchers went further to apply systematic random sampling in choosing mark profiles of fifteen pupils from each of the four selected classes, resulting in using results from sixty pupils in all. All the Geography teachers in the two schools (N = 6) were also involved in this study.

The sample of pupils (N = 60) was considered representative because, according to D.B. Van Dalen (1979), if it is descriptive research (of which the present study is), anything from 10% to 20% of the population is representative enough to warrant generalization of results to the population in question. Sixty pupils are 17% of the population; thus, the sample was representative not only in terms of numbers but also in terms of the kinds of those numbers (pupils with different abilities).

It should be noted that small numbers were preferred because, according to L. Rouse Florian, M. Black-Hawkins and K.S. Jull (2004), statistical analyses that involved very large denominations or samples are usually more likely to show statistical significance even where there is no educational significance at all. For L. Rouse Florian, M. Black-Hawkins and K.S. Jull (2004), reports of significance using large samples should be interpreted and used with caution.

It was for this reason that a small, but representative sample, was preferred in this study. Data was analysed qualitatively and quantitatively, using Person's Product Moment Correlation (r) was used to assess the extent to which the pupils mid-year test scores related to their final examinations scores. A one-way Analysis of Variance (ANOVA) was also conducted to see if there was a significant difference in performance among the four different Geography classes.

CONCEPTUAL FRAMEWORK

Zimbabwe, like many other countries, places heavy emphasis on tests and examinations at all levels of the education system. Although such an education system has been criticized for providing credentials instead of skills to its graduates, there is great unanimity of opinion among many scholars such as E. Mpofu (1991) that teaching without testing is unthinkable. Such a view shows the long-standing tradition of the use of tests to enhance not only teaching and learning but also summative evaluation which, in this case, are final examinations. For H.W. Stevenson, T. Parker and I. Wilkinson (1976), teachers' ratings of pupils' performance play a critical role in pupils' final achievements yet; the usefulness of these ratings has not been adequately studied.

For A. Anastasi (1988) and E. Mpofu (1991), tests are an integral process of teaching and learning which, if properly set and conducted, may enhance quality performance in final examinations. Tests give a picture of where the pupil is and how he/she is likely to perform in the examination. Thus, although the issue of the desirability of tests and examinations is still fraught with controversy; in many countries, tests and examinations are the main means through which teachers and schools give accountability. Tests are and they remain important definers of teaching/learning situations. Through tests, educational aims, objectives, content, methods, and standards are questioned.

What inspired the present writers is that, form both literature and personal experiences, many teacher-made tests lack validity and reliability which are key psychometric properties of any test. Many such tests have been found to have items that are ambiguous, irrelevant, or excessively wordy, yet, they are given such prominence in the education of children.

How valid are these tests? It was against such a backdrop that the present study set

out to investigate the prediction levels of teacher-made tests as a way of assessing their usefulness.

DATA PRESENTATION AND ANALYSIS

Data for all the four classes involved in this study were first presented in tables and briefly interpreted and analyzed before an overall analysis/discussion of the result was done. The tables that follow present the data for this study.

A glance at the results indicates that in the majority of cases, mid-year test scores are generally lower than final examination scores. Assuming that these scores were minimally affected by error of measurement and/or the use of mid-points for final examination symbols, these results may suggest, among other things, that mid-year tests were more difficult than the final examination. Notice that setting tests that are at a wrong or inappropriate difficulty level results in error of measurement. Tables 1 (b) shows how the above scores in table 1 (a) relate to each other.

The results in table 1 (b) indicate a moderate positive relationship (0,560) between mid-year test results and final examination results. This relationship yields statistical significance at the 0,05 level. This means that there is a moderate, positive, and significant relationship between the pupils' scores in the final examination. Therefore, for this class (4A), their mid-year tests scores predict or can be used to predict their performance in the final examinations.

Below is an exploration of results from School Z, Class 4D, lower streamed.

Table 1 (a):
Mid-Year Test Results and Final Examination Results for School Z, Class 4A, Upper Stream (N=15)

Students	Mid-Year Score as %	Final Examination Symbol	Mid-Point
1	57	B	65
2	49	C	55
3	64	C	55
4	71	A	75
5	70	B	65
6	58	A	75
7	73	A	75
8	58	B	65
9	68	C	55
10	69	B	65
11	48	C	55
12	66	A	75
13	67	A	75
14	76	A	75
15	59	B	65

Table 1 (b):
Correlation of Mid-Year Test Scores and Final Examination Scores for School Z, Class 4A, Upper Stream. (N = 15)

		Mid-Year Scores	Final Examination Scores
Mid-year scores	R (r) Sig (2 tailed) N	1,000 15	0,560* 0,030 15
Final Exam scores	R (r) Sig (2 tailed) N	0,560* 0,030 15	1,000 15

Alpha Levels: 0,01 & 0,05
** indicates significance at 0,01 level (2 tailed).
* indicates significance at 0,05 level (2 tailed).

Class 4D is the weakest Form IV class in School Z. All pupils did badly in both the mid-year test and the final examination in Geography. The pupils' scores were correlated and the results are shown in table 2 (b) .

The results in table 2 (b) again show a moderate, positive, and statistically significant relationship (0,591) between mid-year test scores and final examination scores in Geography. This relationship reaches significance at the 0,05 level.

Table 3 (a) shows scores for pupils form Schools Y, Class 4B.

Once again a glance at table 3 (a) shows that in the majority of cases, mid-year test scores are generally lower than final examination scores. As already argued, if this is not the effect of the use of mid-points, then it could be attributed to among other things, error of measurement as in too strict marking by teachers or setting tests that are too difficult.

The scores were correlated to see the extent of their relationship and the result are shown in table 3 (b) .

Table 3 (b) shows a very high positive relationship (0,919) between mid-year test scores and final examination scores. This very high positive relationship which, unlike the relationship in the two classes of School Z which have already been examined, is statistically significant at the 0,01 level.

The scores of Class 4C, School Y, which was the last class in this study, are shown in table 4 (a).

Table 2 (a):
Mid-Year Test Results and Final Examination Results for School Z, Class 4D, Lower Stream (N = 15)

Students	Mid-Year Score as %	Final Examination Symbol	Mid-Point
1	24	U	22
2	22	U	22
3	30	U	22
4	14	U	22
5	38	D	22
6	29	U	47
7	40	E	22
8	37	U	40
9	29	U	22
10	30	U	22
11	20	U	22
12	16	U	22
13	19	U	22
14	33	U	22
15	24	U	22

Table 2 (b):
Correlation of Mid-Year Test Results and Final Examination Results for School Z, Class 4D, Lower Stream (N = 15)

		Mid-Year Scores	Final Examination Scores
Mid-year scores	R (r) Sig (2 tailed) N	1,000	0,591*
			0,020
		15	15
Final Exam scores	R (r) Sig (2 tailed) N	0,591*	1,000
		0,020	
		15	15

Alpha Levels: 0,01 and 0,05
** indicates significance at 0,01 level (2 tailed).
* indicates significance at 0,05 levels (2 tailed).

The pattern that emerges again is that mid-year test scores are generally lower than final examination scores. As already discussed, perhaps this is a result of the use of mid-points or error of measurement.

The above scores were correlated and table 4 (b) shows the results.

Table 4 (b) shows a high positive relationship (0,654) between mid-year test scores and final examination scores. This relationship yields statistical significance at the 0,01 level.

DISCUSSION AND FINDINGS

The results from all the four classes involved in the present study showed a linear positive relationship between mid-year tests scores and final examination scores. The positive relationship, which was significant in all cases and which ranged from moderate to very high, was established in all the classes irrespective of the pupils' abilities. Such findings indicate that teacher-made tests in the Geography classes involved are valid and can, therefore, be reliably used to predict pupils' performance in the final examinations. These findings confirm and extend findings established in other countries. For example, R. Hoste (1981) in Britain and S.A. Glover (1989) in Indiana found that teacher-made tests accurately predicted pupil performance in final "O" level examinations.

Table 3 (a):
Mid-Year Test Results and Final Examination Results for School Y, Class 4B, Mixed Ability Class (N =15).

Students	Mid-Year Score as %	Final Examination Symbol	Mid-Point
1	41	D	47
2	74	A	75
3	20	U	22
4	56	E	40
5	66	A	75
6	61	B	65
7	49	C	55
8	35	U	22
9	37	C	55
10	67	A	75
11	10	U	22
12	51	B	65
13	48	B	65
14	30	U	22
15	45	E	40

Table 3 (b):
Correlation of Mid-Year Test Results and Final Examination Results for School Y, Class 4B, Mixed Ability Class (N =15).

		Mid-Year Scores	Final Examination Scores
Mid-year scores	R (r) Sig (2 tailed) N	1,000	00,919* 0,000 15
Final Exam scores	R (r) Sig (2 tailed) N	0,919* 0,000 15	1,000 15

Alpha Levels: 0,01 and 0,05
** indicates significance at 0,01 level (2 tailed).
* indicates significance at 0,05 levels (2 tailed).

In the present study, interview data from the six Geography teachers indicated that experience in teaching the subject can result in teachers setting tests that are similar to the final examination paper. Other reasons given for pupils' good performance in final examinations include the availability of relevant learning materials such as textbooks, classrooms that limit routinisation in favor of exploratory and discovery learning, frequent testing to ensure consolidation of learned material, and the ability of the teacher to differentiate between what is important and what is not as well as between what is relevant and what is irrelevant when teaching.

Since the two schools involved in this study streamed pupils according to ability, the present researchers became interested in finding out whether the differences in performance observed among the different classes were real (significant) or not. For this, a one-way Analysis of Variance (ANOVA) was applied. The results are shown in table 5.

Results in table 5 indicate that there is a significant difference in performance among the four Geography classes involved in this study (P-Value = 0,000 < 0,05). Therefore, the different classes, which were streamed according to ability, had real differences in their performance. This suggests that upper streams significantly performed between than lower streams. This is perhaps why L.M. Aleamoni (1977) argues that there is bound to be differences among schools and

Table 4 (a):
Mid-Year Tests Results and Final Exam Results for School Y,
Class 4C, Average Class (N = 15)

Students	Mid-Year Score as %	Final Examination Symbol	Mid-Point
1	39	U	22
2	52	B	65
3	49	C	55
4	51	E	40
5	49	C	55
6	50	B	65
7	63	A	75
8	44	U	22
9	51	D	47
10	38	E	40
11	56	E	40
12	52	B	65
13	48	E	40
14	45	C	55
15	45	D	47

Table 4 (b):
Correlation of Mid-Year Test Results and Final Examination Results for School Y,
Class 4C, Average Class (N = 15)

		Mid-year Scores	Final Examination Scores
Mid-year scores	R (r) Sig (2 tailed) N.	1,000 15	0,654* 0,008 15
Final Exam scores	R (r) Sig (2 tailed) N	0,654* 0,008 15	1,000 15

Alpha Levels: 0,01 and 0,05
** indicates significance at 0,01 level (2 tailed).
* indicates significance at 0,05 levels (2 tailed).

Table 5:
One-Way Analysis of Variance for the Four Classes.
Scores (N = 15).

Scores	Sum of Squares	Dif.	Mean Square	F.	Sig.
Between groups	23082,300	3	7694,100	49,045	0,000
Within groups	18198,067	116	156,880		
Total	**41280,367**	**119**			

classes as a function of, among other things, teaching methods, aspects of the sample (different abilities of the different classes), and institutions studied.

Asked whether streaming according to ability was educationally sound, all the six teachers said that it was sound since it allowed teachers to give "appropriate attention" to tend different classes. In the teachers' opinion, streaming ensured that bright pupils were not kept back by slowed learners. This would perhaps explain why many Zimbabwean secondary schools stream pupils according to ability despite teachers' awareness of the negative effects (self fulfilling prophecy) streaming may have on pupils, especially weak pupils.

CONCLUSION

Since positive and significant correlations were found between mid-year test scores and final examination grades in all cases, it can be concluded that to a greater extent, teacher-made tests predict pupils' performance in final examinations. Although the sample in this study was relatively small (N = 60) and although the study was limited to Geography, the result of this study can with caution be extrapolated to other subjects and schools in Zimbabwe.

Basing on the observations made in this study, the following recommendations were made.

First, there is need for more regional and national workshops for all teachers in order to help them not only set valid tests, but also know how to reduce error of measurement when scoring pupils' work. This is critical because, according to E. Mpofu (1991), testing is the core of teaching and learning; teaching without testing is unthinkable.

Second, colleges and universities should ensure that intending teachers are well grounded in measurement theory. From the writers' knowledge and experience, little is done in this area by colleges and universities.

Third, finally, perhaps there is need for a larger study involving many schools and subjects to see if the results of the present study can be confirmed or even extended.

Bibliography

Aleamoni, L.M. (1977). "Can Grade Point Average be More Accurately Predicted?" in *Journal of Educational Psychology,* 69, pp.227-230.

Anastasi, A. (1988). *Psychological Testing*. New York: MacMillan.

Florian, L. Rouse, M. Black-Hawkins & K.S. Jull. (2004). "What Can National Data Sets Tell Us about Inclusion and Pupil Achievement?" in *British Journal of Special Education*, 31(3), pp.115-121.

Glover, S.A. (1989). "The Testing Phenomenon, Not Gone by Nearly Forgotten" in *Journal of Educational Psychology,* 76, pp.777-781.

Hoste, R. (1981). "How Valid are School Examinations? An Exploration into Content Validity" in *British Journal of Educational Psychology*, 51, pp.10-22.

Kerlinger, F.N. (1986). *Foundations of Behavioral Research*. New York: Harcourt.

Mehrens, W.A. & I.J. Lehmann. (1978). *Measurement and Evaluation in Education and Psychology*. New York: Holt.

Mpofu, E. (1991). *Testing for Teaching*. Harare, Zimbabwe: Longman.

Mwatengahama, N. (1999). "A Comparison of Teaching Rating and Standardized Test Scores in Reading and Mathematics for Slow Learners at Secondary School Level". *Unpublished Med Dissertation*. Zimbabwe: Department of Educational Foundations, University of Zimbabwe.

Stevenson, H.W., T. Parker & I. Wilkinson. (1976). "Predictive Value of Teachers' Ratings of Young Children" in *British Journal of Educational Psychology*, 68, pp.507-517.

Van Dalen, D.B. (1979). *Understanding Educational Research: An Introduction*. New York: Holt.

www.google/maps.com [accessed in Harare, Zimbabwe: 21 April 2013].