PN-ABP-390
ISN 82927

RESEARCH PAPERS

BOTSWANA

Research and Testing Centre (RTC)
Curriculum Development and Evaluation

Ministry of Education
RTC RESEARCH NOTES

March 1988

IEES

Improving the Efficiency of Educational Systems

Florida State University
Howard University
Institute for International Research
State University of New York at Albany

United States Agency for International Development
Bureau for Science and Technology
Office of Education
Contract No. DPE-5283-C-00-4013-00

Improving the Efficiency of Educational Systems (IEES) is an initiative funded in 1984 by the Agency for International Development (AID), Bureau for Science and Technology, Office of Education. The principal goals of the IEES project are to help developing countries improve the performance of their educational systems and strengthen their capabilities for educational planning, management, and research. To achieve these goals, a consortium of U.S. institutions has been formed to work collaboratively with selected host governments and USAID Missions for ten years. The consortium consists of The Florida State University (prime contractor), Howard University, the Institute for International Research, and the State University of New York at Albany.

There are seven countries working with the IEES initiative to improve educational efficiency: Botswana, Haiti, Indonesia, Liberia, Nepal, Somalia, and Yemen Arab Republic.

Documents published by IEES are produced to promote improved educational practice, planning, and research within these countries. All publications generated by project activities are held in the IEES Educational Efficiency Clearinghouse at The Florida State University. Requests for project documents should be addressed to:

IEES
Educational Efficiency Clearinghouse
Learning Systems Institute
204 Dodd Hall
The Florida State University
Tallahassee, Florida 32306
USA
(904) 644-5442

Agency for International Development
Bureau for Science and Technology
Office of Education
Contract No. DPE-5823-C-00-4013-00
Project No. 936-5823

PN-ABP-390

BOTSWANA

Research and Testing Centre (RTC)
Curriculum Development and Evaluation

Ministry of Education

RTC RESEARCH NOTES

Serara Moahi
John Bowers

March 1988

Table of Contents

Note 88-01 March 1988
A Comparison of T-Scores and Normalized T-Scores for Reporting Performance on the 1987 Setswana Primary School Leavers Examination (PSLE) 1

Note 88-02 April 1988
Comparing T-Scores and Normalized T/N-Scores for Paper 2 of the 1987 Mathematics PSLE 3

Note 88-03 April 1988
A Statistical Analysis of the 1987 Two-Year Social Studies Junior Certificate Examination 7

Note 88-04 June 1988
An Analysis of Results for the 1987 Two-Year Junior Certificate Examination in Art 18

Note 88-05 July 1988
A Preliminary Analysis: The 1987 Administration of the IEA Tests to Two-Year and Three-Year JC Completers 25

Note 88-06 July 1988
1987 Two-Year JC Multiple-Choice Item and Total Score Statistics 34

Note 88-07 July 1988
Combining JC Within-Subject Scores: Nominal and Effective Weighting 47

Note 88-08 July 1988
JCE Overall Merit Ratings Determined from the Average of Seven Subject Standard Scores 51

Note 88-09 July 1988
A Frame for Writing Examination Items 54

A Comparison of T-Scores and Normalized T-Scores for Reporting Performance on the 1987 Setswana Primary School Leavers Examination (PSLE)

Research Note 88-01

March 1988

Research and Testing Centre
Curriculum Development and Evaluation
Ministry of Education
Government of Botswana

A Comparison of T-Scores and Normalized T-Scores for Reporting Performance on the 1987 Setswana PSLE

Problem

T-Score distributions for both papers of the 1987 Setswana PSLE show a large negative skew. Scores tend to pile up at the high end of the marking scale. This results in a smaller proportion of examinees achieving higher T-Scores and letter grades than would be expected with symmetric distributions.

This point is illustrated in Figure 1 which shows the distribution of T-Scores for the multiple choice paper of the 1987 Setswana PSLE. The raw score mean is 39.9 with a standard deviation of 9.3. This is linearly transformed to a T-Score with a mean of 50 and a standard deviation of 10. Since the maximum mark is 60, the maximum T-score is just above 70.

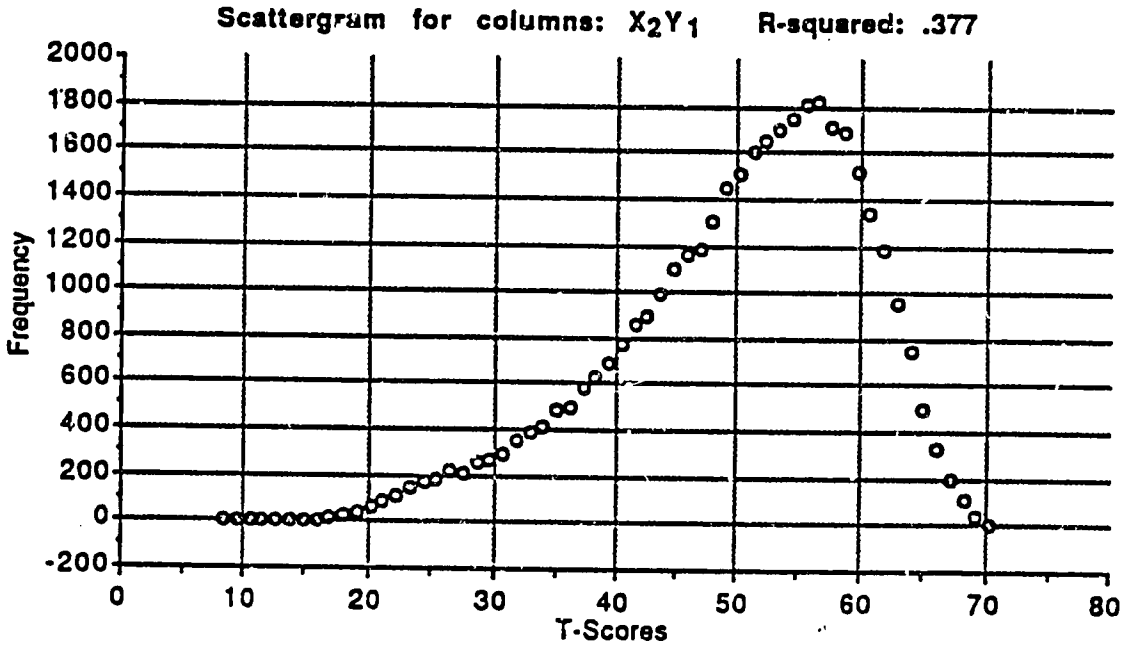

Figure 1. Distribution of T-Scores: 1987 Setswana PSLE Multiple Choice Test

For purposes of discussion, consider that a letter grade of "A" is assigned to any respondent achieving a T-Score of 63 or higher. For the distribution shown in Figure 1, an "A" would be assigned to 7.6 per cent of the respondents. In comparison, one would expect that 9.7 per cent of respondents would achieve T-Scores of 63 or higher in a normal distribution. The difference between 9.7 per cent and 7.6 per cent may seem small, but in a group of nearly 40, 000 examinees, the difference is one between approximately 3900 "A" grades and 3000 "A" grades.

Normalized T-Scores (T/N-Scores)

Skewness can be removed by transforming raw scores to Normalized T-Scores (T/N-Scores). These are also known as McCall's T-Scores. In a distribution of T/N-Scores, the proportion of examinees at or above each T/N-Score value is the same as that for the normal distribution. The distribution of T/N-Scores for the multiple choice paper of the 1987 Setswana PSLE is shown in Figure 2.

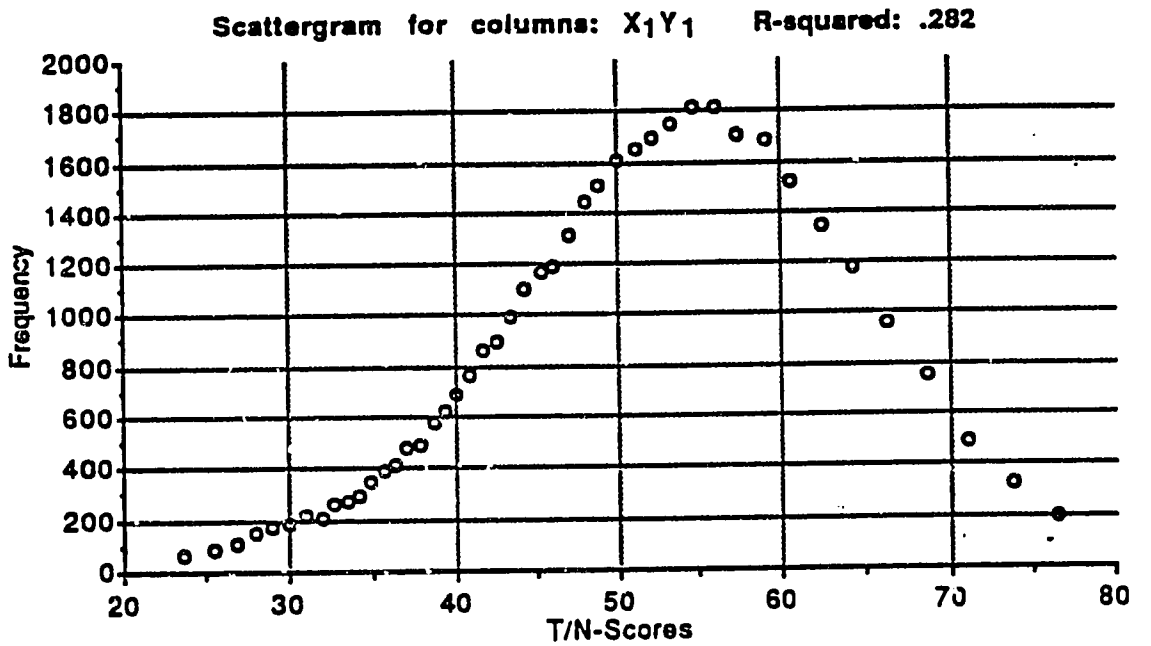

Figure 2. Distribution of T/N-Scores: 1987 Setswana PSLE Multiple Choice Test

The distribution in Figure 2 approximates the familiar bell-shape of the normal distribution. Since T/N-Score values extend to higher levels than do the T-Score values shown in Figure 1, it follows that a larger proportion of respondents will achieve higher T/N-Scores than T-Scores. In the example above with a T-Score of 63 as the cutoff point for "A" grades, 10.7 per cent of the examinees achieve a normalized T/N-Score of 63.

Recommendation

It is recommended that PSLE policy base letter grades in Setswana upon T/N-Scores rather than T-Scores. The following factors bear upon this consideration.

- T/N-Score transformation enables a full range of letter grades for PSLE marks.
- T/N-Score percentile ranks are the same as normal curve percentile ranks so can be specified.
- T/N-Scores have the same mean and standard deviation as T-Scores now used, so no problem should arise for school administrators and teachers who have become familiar with the scale.
- T/N-Score transformation is easily computerized, requiring only a normal curve look up table.

This procedure should perhaps also be considered for other PSLE subjects where raw score distributions are asymmetric.

Comparing T-Scores and Normalized T/N-Scores for Paper 2 of the 1987 Mathematics PSLE

Research Note 88-02

April 1988

Research and Testing Centre
Curriculum Development and Evaluation
Ministry of Education
Government of Botswana

Comparing T-Scores and Normalized T/N-Scores for
Paper 2 of the 1987 Mathematics PSLE

A preceding note (Research Note 88-01) compared linearly transformed T-Scores with normalized T/N-Scores for reporting performance on the multiple choice paper of the 1987 Setswana PSLE. It was recommended that PSLE letter grades be based upon T/N-Scores when distributions are found to be markedly skewed because they enable percentile ranks and the percentage of letter grades to be predictably specified in terms of the normal curve. It was suggested that a switch to T/N-Score reporting be considered when the raw score distributions for other subject examinations are asymmetric.

An example arises for Paper 2 of the 1987 Mathematics PSLE. The T-Score distribution for this examination is shown in Figure 1. Its high degree of negative skew constrains the achievement of higher T-Scores and associated letter grades. The distribution of T/N-Scores for the 1987 Mathematics PSLE is shown in Figure 2. T/N-Scores range to a high of 80 compared to a maximum T-Score of 65.

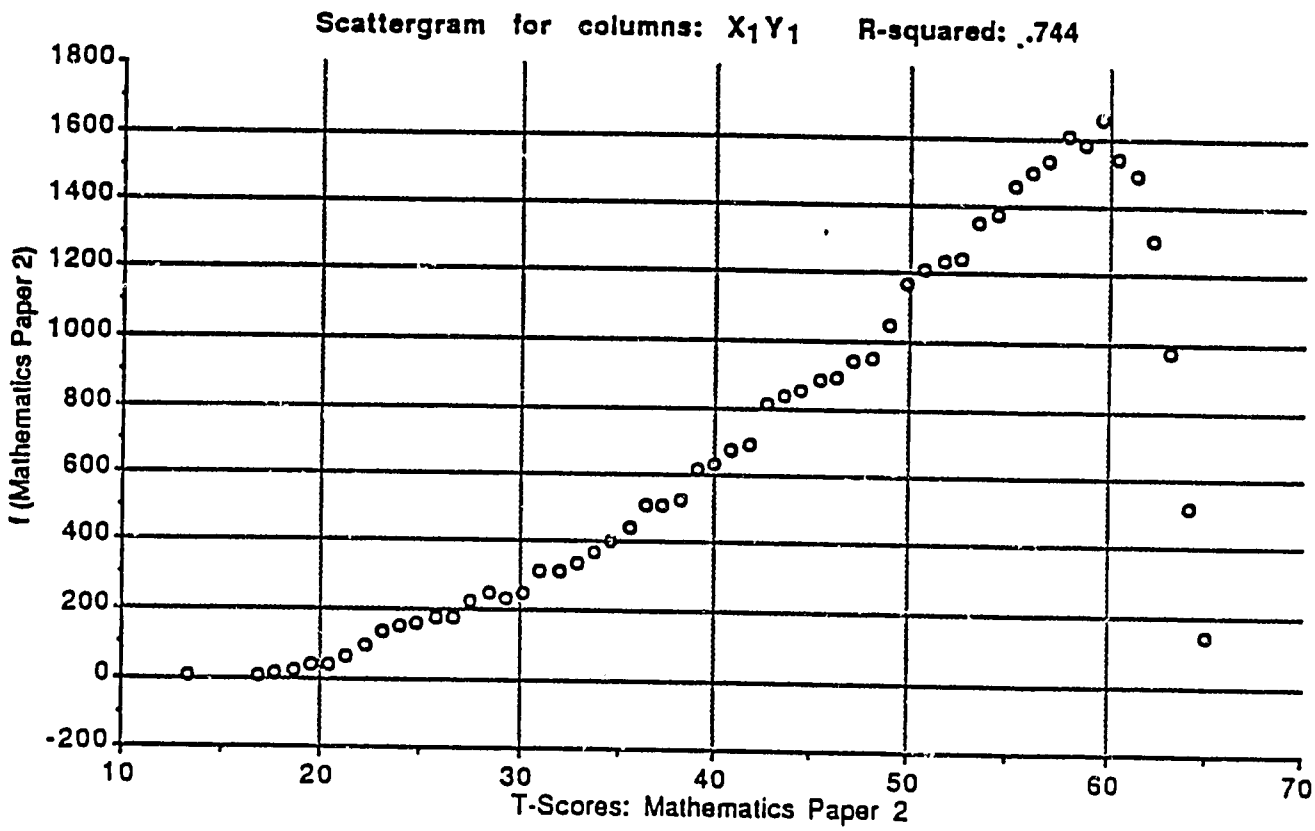

Figure 1. Distribution of T-Scores: 1987 Mathematics PSLE Paper 2.

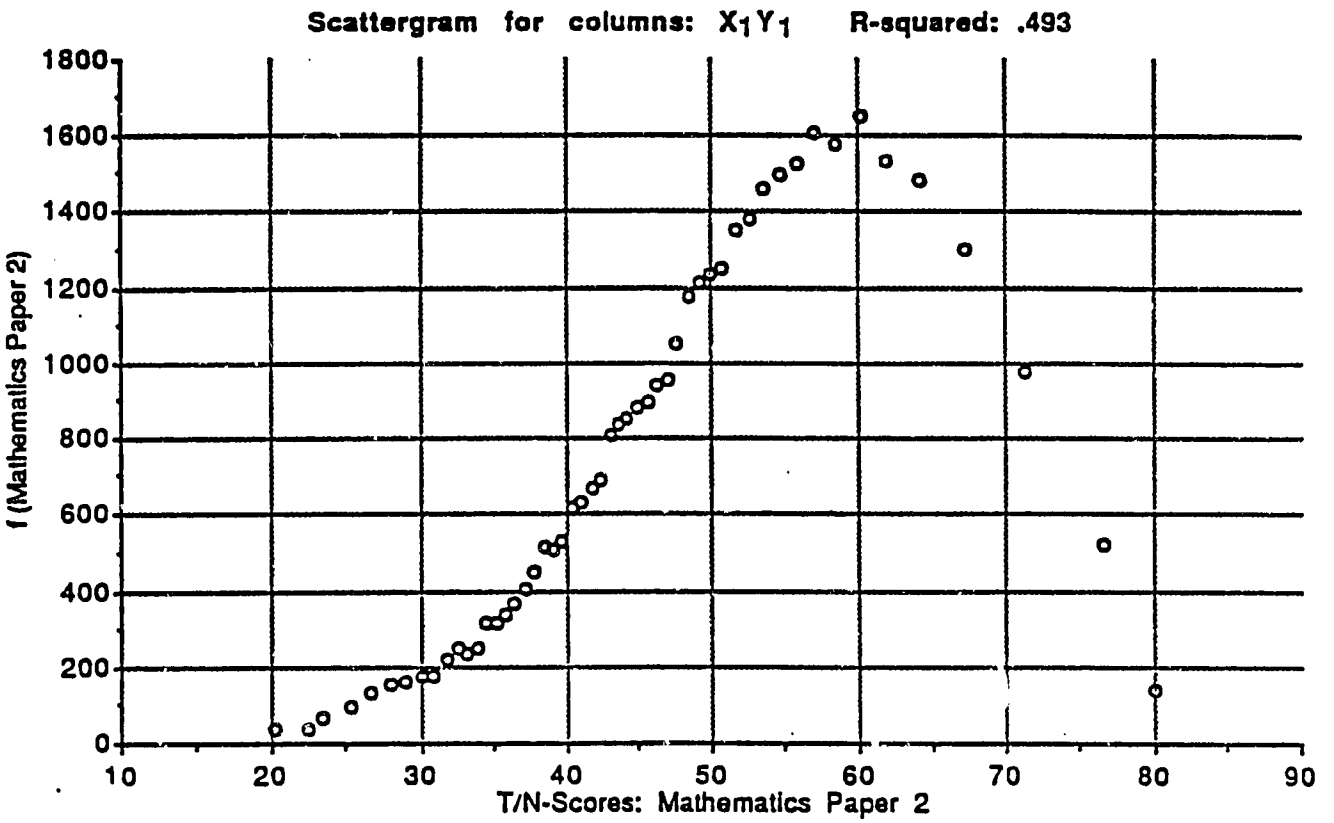

Figure 2. Distribution of T/N-Scores: 1987 Mathematics PSLE Paper 2.

Table 1 shows the percentage of examinees who achieve comparable T-Scores and T/N-Scores at the high end of the scale. For the sake of illustration, if a value of 63 represents an "A" letter grade, then 1.7 per cent of the examinees achieve an "A" with T-Score transformation and 9.2 per cent with T/N-Score transformation.

Table 1

A Comparison of T-Scores and T/N-Scores

Score	Percentage Achieving With T/N-Scores	T-Scores
63 or higher	9.2 %	1.7 %
62 or higher	11.5	4.2
61 or higher	13.4	11.5
60 or higher	15.5	15.5

For this highly skewed test, the T/N-Score transformation leads to more desirable marking in the higher ability range than does the T-Score transformation. Nevertheless, it needs to be pointed out that transforming raw scores to normalized T/N-Scores treats the **effects** of skewness but not the **cause**. The T/N-Score transformation provides an answer but not a solution to the problem.

Skewness is best controlled by selecting items with known statistical characteristics. The relationship between the mean and the standard deviation of the total test score distribution and the familiar item statistics p_i (the facility for the i-th item) and r_{it} (the point biserial correlation between the i-th item score and the total test score) is:

$$\text{Test mean} = \sum_i p_i$$

$$\text{Test standard deviation} = \sum_i r_{it}\sqrt{p_i(1-p_i)}$$

Pretesting permits the estimation of p_i and r_{it}. Items can be selected to produce a total test score distribution with a mean about three standard deviations below the maximum test score. T-Scores would then have a maximum value of about 80 and marking at the high end of the scale would not be attenuated.

A Statistical Analysis of the 1987 Two-Year Social Studies Junior Certificate Examination

Research Note 88-03

April 1988

Research and Testing Centre
Curriculum Development and Evaluation
Ministry of Education
Government of Botswana

A STATISTICAL ANALYSIS OF THE 1987 TWO-YEAR SOCIAL STUDIES JUNIOR CERTIFICATE EXAMINATION

INTRODUCTION

This report presents a statistical analysis of the Botswana 1987 Two-Year Junior Certificate Examination (JCE) in social studies. The summary data has been prepared to help staff of RTC and members of the social studies examinations group to improve future JCE tests.

THE 1987 TWO-YEAR SOCIAL STUDIES JCE

Paper 1 presents 50 four-option multiple-choice items. One mark is given for each correctly answered item, with no correction for guessing. Paper 1's total score is number correct. Testing time is one hour.

Paper 2 presents four interpretive exercises with questions based upon tabled data, graphic information, a map of Southern Africa, and a short reading passage. Maximum mark is 10 for each exercise. Testing time is one hour.

Paper 3 is an extended response examination that presents six topics. Respondents are asked to write from one to one and one-half pages on three of these. Maximum mark is 20 for each topic. Testing time is one and one-half hours.

ANALYSIS

The analysis reports the following information.

- Raw score distributions, means, and standard deviations for the three separate papers;
- Item statistics within each paper;
- The distribution and statistics for the total score -- the sum of the three papers' raw scores;
- Intercorrelations among the scores for separate papers and the total score;
- A comparison of intended and actual weighting of the separate paper raw scores;
- Reliability estimates for all scores.

Except for Paper 1's item analysis which is based upon data for all respondents in 1987, the analysis is otherwise based upon a sample of 312 students in six representative JCE centres.

RAW SCORE DISTRIBUTIONS FOR SEPARATE PAPER SCORES

Figures 1 through 3 present histograms for the separate paper scores. The mean, standard deviation, and maximum mark are reported for each.

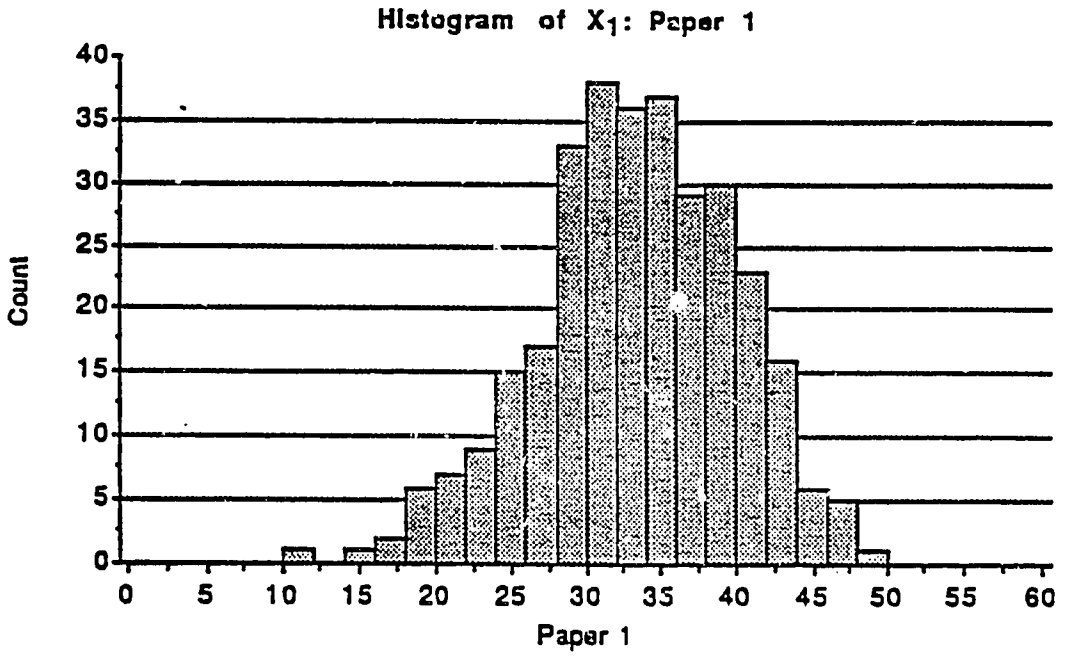

Fig. 1. Histogram for 1987 Two Year Social Studies JCE Paper 1.
(Mean= 32.9; S.D.=6.5; Maximum mark=50)

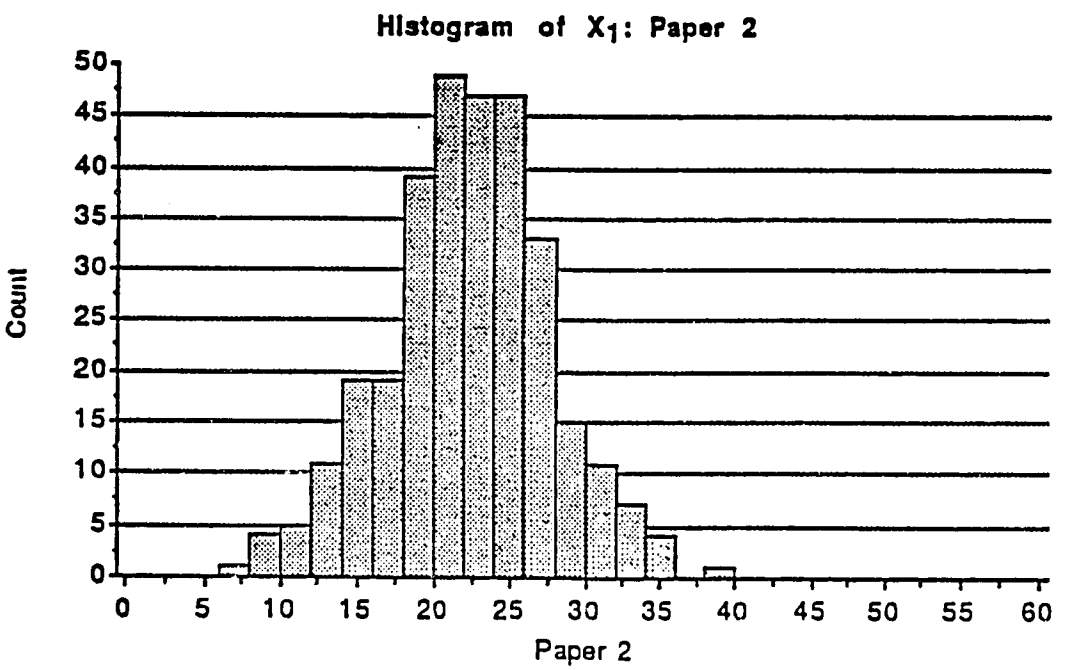

Fig. 2. Histogram for 1987 Two Year Social Studies JCE Paper 2.
(Mean= 21.7; S.D.= 5.3; Maximum mark=40)

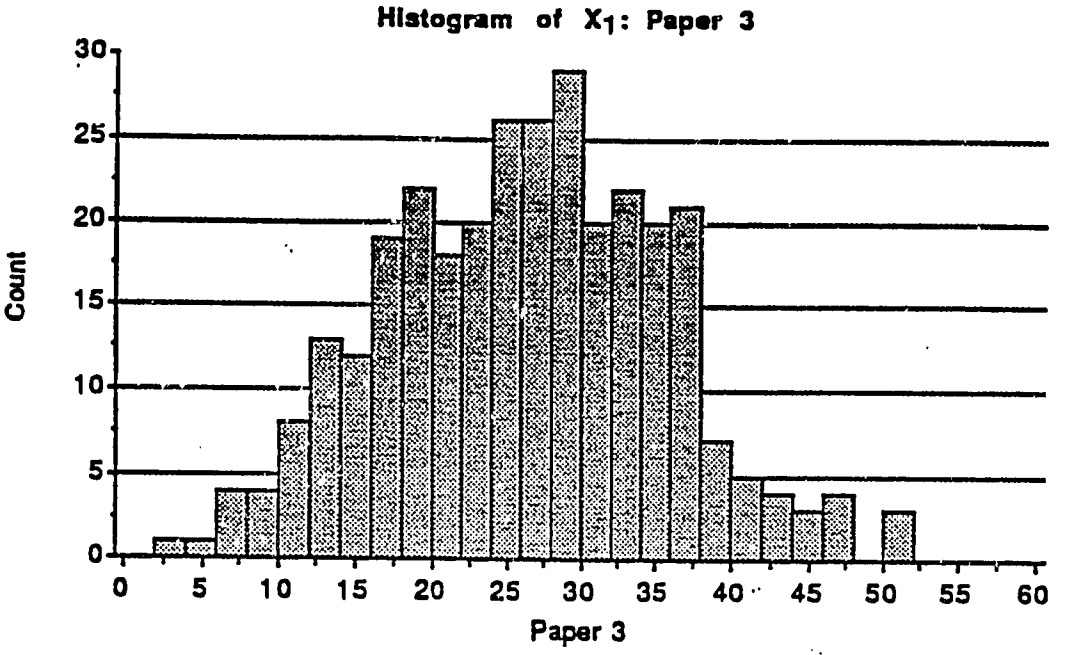

Fig. 3. Histogram for 1987 Two Year Social Studies JCE Paper 3.
(Mean= 25.7; S.D.= 9.3; Maximum mark=60)

ITEM STATISTICS FOR SEPARATE PAPERS

Paper 1. Table 1 lists item analysis statistics for the multiple-choice Paper 1. Statistics reported are for all 1987 respondents. Items are scored 1 if correctly answered, 0 otherwise. Total score is number correct; the maximum score is 50.

- **Item p** is the proportion of respondents who answered each item in Paper 1 correctly;
- **Item r** is the correlation of the item score with the total score for Paper 1;
- **Item S** is the item score's contribution to the total score's standard deviation.

Table 1

Item Statistics for Paper 1 of the 1987 Two-Year JCE in Social Studies

Item No.	Item p	Item r	Item S	Item No.	Item p	Item r	Item S
1	.75	.38	.16	26	.95	.17	.04
2	.76	.30	.13	27	.51	.38	.19
3	.20	.19	.08	28	.61	.09	.04
4	.55	.50	.25	29	.45	.36	.18
5	.89	.28	.09	30	.42	.30	.15
6	.88	.32	.10	31	.65	.05	.02
7	.77	.29	.12	32	.88	.20	.06
8	.78	.27	.11	33	.37	.13	.06
9	.94	.24	.06	34	.38	.42	.20
10	.60	.33	.16	35	.87	.24	.08
11	.38	.06	.03	36	.48	.32	.16
12	.91	.26	.07	37	.90	.25	.07
13	.63	.35	.17	38	.58	.31	.15
14	.36	.33	.16	39	.96	.23	.05
15	.87	.35	.12	40	.53	.16	.08
16	.44	.42	.21	41	.89	.21	.07
17	.86	.26	.09	42	.65	.38	.18
18	.79	.43	.18	43	.43	.35	.17
19	.33	.18	.08	44	.41	.20	.10
20	.96	.17	.03	45	.65	.43	.21
21	.64	.33	.16	46	.43	.39	.19
22	.75	.30	.13	47	.73	.38	.17
23	.57	.39	.19	48	.61	.32	.16
24	.88	.21	.07	49	.58	.34	.17
25	.78	.10	.04	50	.32	.38	.18
Totals					32.5		6.1

Item p values average 0. 67 and sum to 32.5 which is the total score mean. Item S values sum to 6.1 which is the total score standard deviation. These values are for all students who sat the examination and correspond closely to the respective values of 32.9 and 6.5 for Paper 1 based upon the sample of 312 examinees reported in Figure 1.

Figure 4 shows a scatterplot of Item p and Item S. In Figure 4, 20 items have Item S values below 0.10. These 20 items contribute an average of 0.06 and a combined total of 1.2 (20 per cent) to Paper 1's standard deviation of 6.1; 12 of the 20 items also have Item p values above 0.80, which is too many easy items for a 50-item test. These twenty items will be stored for future revision.

Thirty items functioned well--those with Item p values between 0.30 and 0.80 and Item S values of 0.10 or higher. These will be item banked. The 30 items contributed an average contribution of 0.16 and a combined contribution of 4.9 to the total score standard deviation of Paper 1. A test comprised of 50 items with Item p and Item S values like those for these 30 items would have a mean total score of approximately 29.6 and a total score standard deviation of 8.2. This hypothetical test would discriminate better across a wider range of ability than the 1987 Social Studies JCE Paper 1.

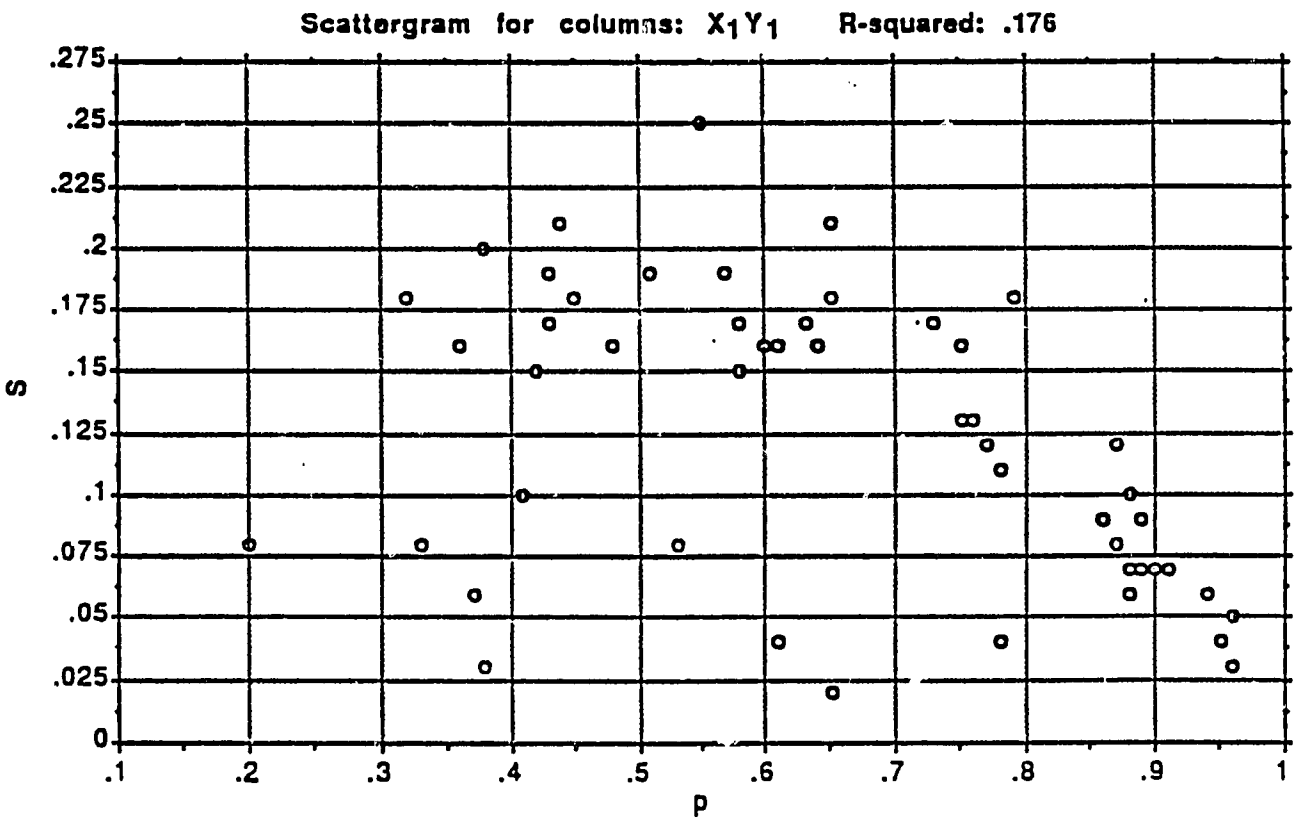

Fig. 4. Scatterplot of Item p and Item S for Items on Paper 1 (taken from Table 1).

Paper 2. Table 2 presents statistics for each of the four interpretive exercises in Paper 2. Each exercise has a maximum score of 10; the maximum total score is 40. The analysis is based upon the sample of 312 respondents in six representative centres.

- **Item Mean** is the average score for each exercise;
- **Item SD** is the standard deviation of the scores for each exercise;
- **Item r** is the correlation between each exercise score and the total score for Paper 2;
- **Item S** is the contribution of each exercise score to the standard deviation of the total score.

Paper 2 requires the application of interpretive skills tested with four kinds of presented material; exercises are based upon a table, a bar chart, a map, and a reading passage. Table 2 indicates large mean differences in average marks given the four exercises as well as differences in the contribution each exercise makes to Paper 2's total score standard deviation.

Table 2

Item Statistics for Paper 2 of the1987 Two-Year JCE in Social Studies (N=312)

Exercise Number	Item Mean	Item SD	Item r	Item S	Intercorrelations 1	2	3	4
1	7.7	1.7	.61	1.1	1.00			
2	3.1	1.9	.64	1.2	.19	1.00		
3	6.5	2.2	.66	1.5	.22	.25	1.00	
4	4.4	2.2	.67	1.5	.27	.22	.19	1.00
Total	21.7	5.3		5.3	.61	.63	.67	.67

Item M sums to the total score mean; Item S sums to the total score standard deviation.

Paper 2 functions well except for Exercise Two, which is based upon the interpretation of a bar graph. The mean mark for Paper 2 is low and probably due to difficulty in one part of the exercise worth 5 marks that requires respondents to construct a pie chart corresponding to the presented bar graph. This kind of interpretive transformation may be better tested in mathematics.

Intercorrelations among the marks for the four exercises are uniformly low, ranging from 0.19 to 0.27. These are short answer exercises that are objectively marked; they appear to measure different specific interpretive skills depending upon the kind of material to be interpreted.

Paper 3. Table 3 shows statistics for scores on each of the six extended response topics in Paper 3. Data are based upon the responses of the sample of 312 respondents. Respondents select three topics out of six choices. Each of the topics has a maximum score of 20; the maximum total score is 60.

- **N and % Choosing** is the number and percentage of respondents who chose each topic;
- **Item Mean** is the mean score for each topic;
- **Item SD** is the standard deviation of the scores for each topic;
- **Item r** is the correlation between each topic score and the total for Paper 3.

There are large differences in average marks, so choice of a topic appears related to how well a student performs on Paper 3. The percentages of respondents selecting each topic ranged from 85 per cent for Topic 1 to 7 per cent for Topic 6. In effect, different respondents take different tests. Test constructors could consider reducing respondents' choice of topics. Uniformity of topics would fairer individual comparisons of performance. The argument against reducing choice is that schools now vary in subject matter coverage and emphasis, so respondents should be allowed to select topics to better test fairness, at least until new curriculum materials have been implemented. A compromise would be a mix of compulsary and self-selected topics.

Table 3

Item Statistics for Paper 3 of the 1987
Two-Year JCE in Social Studies (N=312)

Topic Number	N and % Choosing		Item Mean	Item SD	Item r
1	265	85%	14.2	5.0	0.73
2	59	19	5.5	4.1	0.60
3	141	45	4.6	3.7	0.55
4	250	80	7.1	3.6	0.63
5	200	64	7.0	4.1	0.61
6	21	07	5.6	4.8	0.69
Total Score	312		25.7	9.3	–

DISTRIBUTION OF THE SUM OF THE THREE SEPARATE PAPER SCORES

Figure 5 presents the distribution, mean, and standard deviation of the sum of the three paper raw scores. The summed score distribution for the total population of respondents is the basis for transforming summed scores to T-Scores with a mean of 50 and a standard deviation of 15. T-Scores are the basis for the assignment of letter grades.

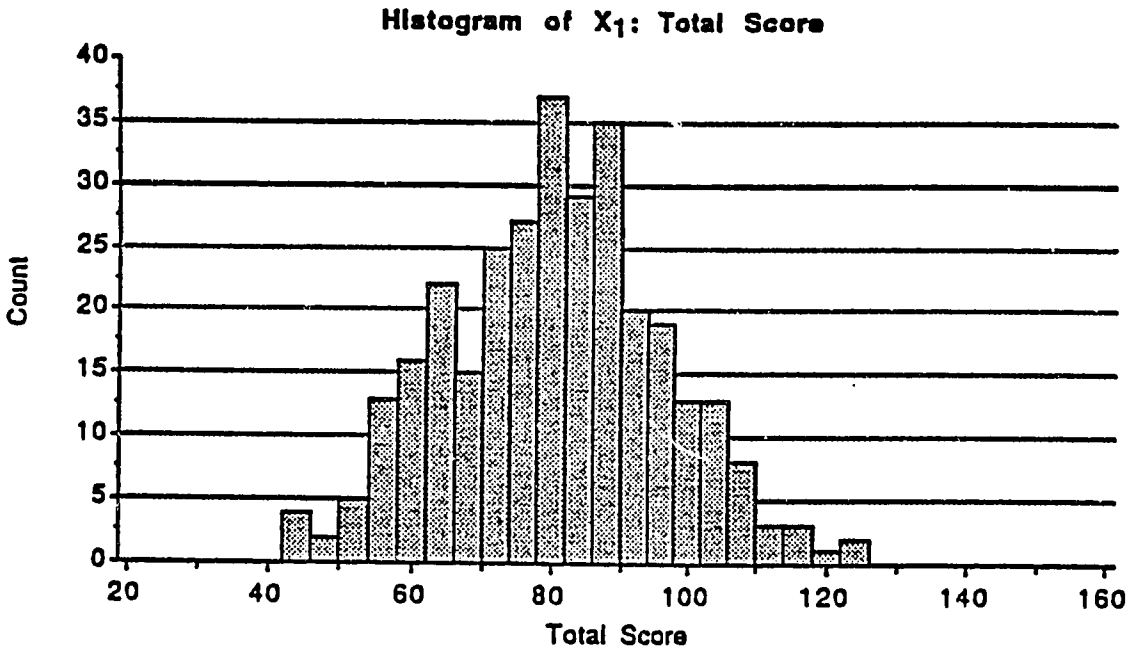

Fig. 5. Histogram for the Sum of the Scores on 1987 Two Year Social Studies JCE Papers 1, 2, and 3. (Mean= 80.3; S.D.= 15.6; Maximum mark=150)

INTERCORRELATIONS AMONG SCORES

Table 4 presents means, standard deviations, and intercorrelations for the separate paper raw scores and the summed total score. Scores for all three papers are moderately correlated, but the correlation between Paper 1 and Paper 3 is somewhat larger than the the correlation of either with Paper 2.

Table 4

Means, Standard Deviations, and Intercorrelations for 1987 Two Year Social Studies Raw Scores (N=312)

Raw Score	Mean	S.D.	Intercorrelations Paper 1	Paper 2	Paper 3
Paper 1	32.9	6.5	1.00		
Paper 2	21.7	5.3	.31	1.00	
Paper 3	25.7	9.3	.40	.19	1.00
Total Score	80.3	15.6	.76	.58	.82

THE WEIGHTING OF SEPARATE PAPER SCORES

When the separate paper scores are added, what is the actual weighting of the separate scores compared to the intended weighting? Actual weights are determined by the statistical properties of the separate raw scores that are summed.

Intended weights for the separate paper raw scores are inferred from the relative number of maximum marks assigned each paper. For the Two Year 1987 Social Studies JCE, assigned marks were 50 for Paper 1, 40 for Paper 2, and 60 for Paper 3. It is assumed that the test developers desired intended weights in the respective ratio of 50 : 40 : 60.

The actual weights of the separate paper raw scores were determined as the contribution of each paper's score to the total score standard deviation; the respective weights were in the ratio of 47 : 30 : 73. The comparison is shown in Table 5. Paper 1's actual weight is slightly less than intended, Paper 2 is weighted considerably less than intended, and Paper 3 considerably more than intended.

Table 5

A Comparison of Intended and Actual Weightings of the Separate Papers

Paper	Intended Weights		Actual Weights	
Paper 1	50	33%	47	31%
Paper 2	40	27	30	20
Paper 3	60	40	73	49
Totals	150	100%	150	100%

THE RELIABILITY OF SCORES

Table 6 presents the reliability estimate (coefficient alpha) for each of the separate paper scores and the total score. Coefficient alpha is a lower bound to the proportion of variance measured in common by items in a test. To the extent that each paper measures specific abilities as well as common abilities also measured by the other papers, the reliability for each paper is higher than its estimated coefficient alpha. Nevertheless, test developers would wish to make each test, whatever factors it measures, more reliable. The important point is that when a single score is used to summarize test performance, high reliability supports sounder interpretation of test scores. If estimates of reliability for a test are low, then inspection of the statistical characteristics of the items in the test may suggest ways to improve its reliability.

Table 6

Coefficient Alpha Estimates

Paper	Alpha
Paper 1	0.70
Paper 2	0.53
Paper 3	0.46
Total Score	0.54

CONCLUSION

The conclusion will briefly summarize a few points with regard to developing future JCE examinations.

Analysis of Testing Results. Routine post mortem statistical analysis of JCE results is needed both at the score and the item levels.

Analysis at the individual paper's total score level provides evidence for examining the relative contribution of each paper to the summed total score for each examination. The summed total score is the basis for transformation to T-Scores and for the assignment of letter grades that reflect overall performance.

Analysis at the individual item level within a paper provides evidence for examining the difficulty level and discrimination of each item. Improved item discrimination enhances the reliability and interpretation of scores over wide ranges of ability.

Item Banking. The need is to pre-specify the statistical properties of JCE examinations. Item p and Item S values for items within a test directly relate to the test's total score mean and standard deviation. When the statistical properties of tests can be specified in advance, then equivalent test papers in a subject can be constructed from one year to the next.

Item banks are the means to develop statistical control. When sufficient numbers of items are available, then test developers have the capability to select items with desired statistical characteristics that lead to predictable score distributions.

Item banks consist of two kinds of items: (a) those that have appeared in live examinations and that have showed desirable statistical properties; (b) pretested items that are statistically eligible candidates for inclusion in future live examinations. Items used in live tests, analyzed, and banked will need to be held out for a time, and somewhat modified for re-use. Pretesting is essential to produce large numbers of bankable items.

Test Development Content and Statistical Specifications. Desirable statistical properties are necessary but not sufficient criteria for good examination items. Items must first of all reflect the content to be covered by an examination and the levels of learning to be measured. A well-defined frame for item preparation must be developed and followed. The social studies JCE developers did this for the 1987 Two-Year Social Studies JCE. Items were prepared to cover specific units tested, and to measure four levels of learning -- basic facts and knowledge, ideas and concepts, applications, and higher level skills and abilities. Items prepared for the 1988 Two-Year Social Studies JCE followed these same specifications and items prepared so far have been pre-tested.

An Analysis Of Results For The 1987 Two-Year Junior Certificate Examination In Art

Research Note 88-04

June 1988

Research and Testing Centre
Curriculum Development and Evaluation
Ministry of Education
Government of Botswana

An Analysis Of Results For The 1987 Two-Year Junior Certificate Examination In Art

Introduction

This report presents an analysis of the score results for the Botswana 1987 Two-Year Junior Certificate Examination (JCE) in art. This was the first year in which the Two-Year JCE in art was administered. Reported JCE results are for 258 examinees in seven schools where art was introduced as an examinable option in 1986.

The 1987 Two-Year Art JCE

JCE performance is based upon total score on a multiple-choice paper and assessment ratings of five pieces of artwork submitted in a portfolio.

Paper 1 presents 40 four-option multiple-choice items that largely cover elements and principles of art, but also includes items relating to art heritage and appreciation. Items in Paper 1 are scored 1 if answered correctly, 0 otherwise. The score for Paper 1 is number of items answered correctly. Maximum score is 40. Testing time is one hour.

Art Portfolio. The Art Portofolio for each examinee contains five pieces of artwork completed during his or her two-year junior secondary art course. Four pieces -- one each in painting, drawing, sculpting, and crafts -- are selected by the art teacher to represent the student's best work. The fifth piece is selected by the student with guidance from the teacher. Three raters evaluated the portfolio submissions on five criteria -- use of visual elements, visual stimulation, use of media, craftsmanship, and one criterion specific to the fifth selected piece. Maximum rating score for each artwork piece on each of the five criteria is 22. The maximum Portfolio score is 110.

Analysis

The following information is presented.

- Histograms for the multiple-choice Paper 1 score, for the Portfolio rating score, and for the Total score which is the sum of the Paper 1 and the Portfolio scores.
- Intercorrelations among the Paper 1, Portfolio, and Total scores.
- Item statistics for the multiple-choice Paper 1.
- The comparative mean achievement in seven schools which introduced art in 1986.

Score Histograms and Statistics

Score Histograms. Histograms for the Paper 1 score, the Portfolio score, and the Total score are shown in Figures 1, 2, and 3. Distributions for all scores are negatively skewed which indicates that students scored moderately high on both the multiple-choice Paper 1 and the Portfolio. The average score of 28.1 for Paper 1 is approximately 70 per cent of the maximum score of 40; the average score of 64.8 on the Portfolio is equivalent to 60 per cent of the maximu score of 110; and the average Total score of 92.9 is 62 per cent of the maximum score of 150.

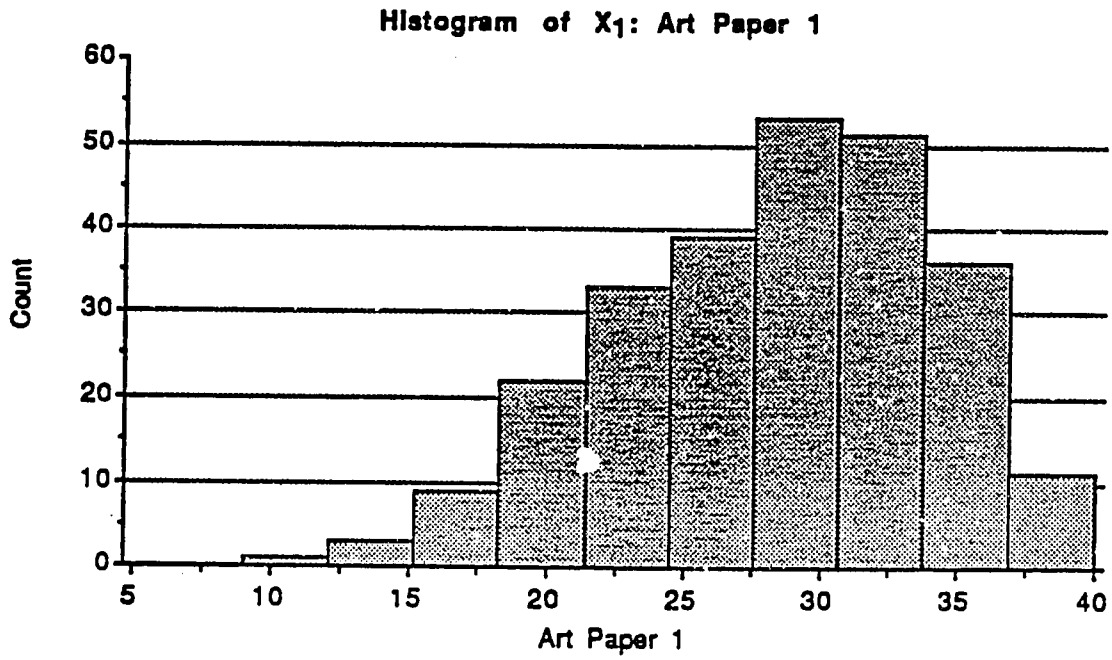

Figure 1. 1987 Two-Year Art JCE Paper 1 Score Distribution: Mean = 28.1

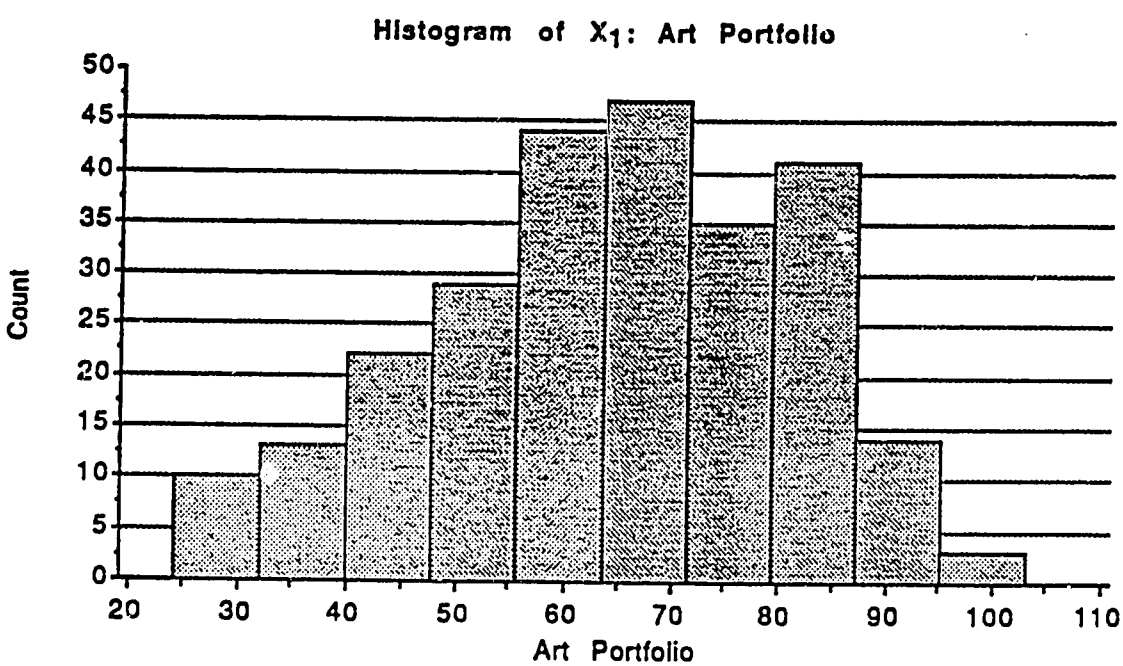

Figure 2. 1987 Two-Year JCE Art Portfolio Score Distribution: Mean = 64.8

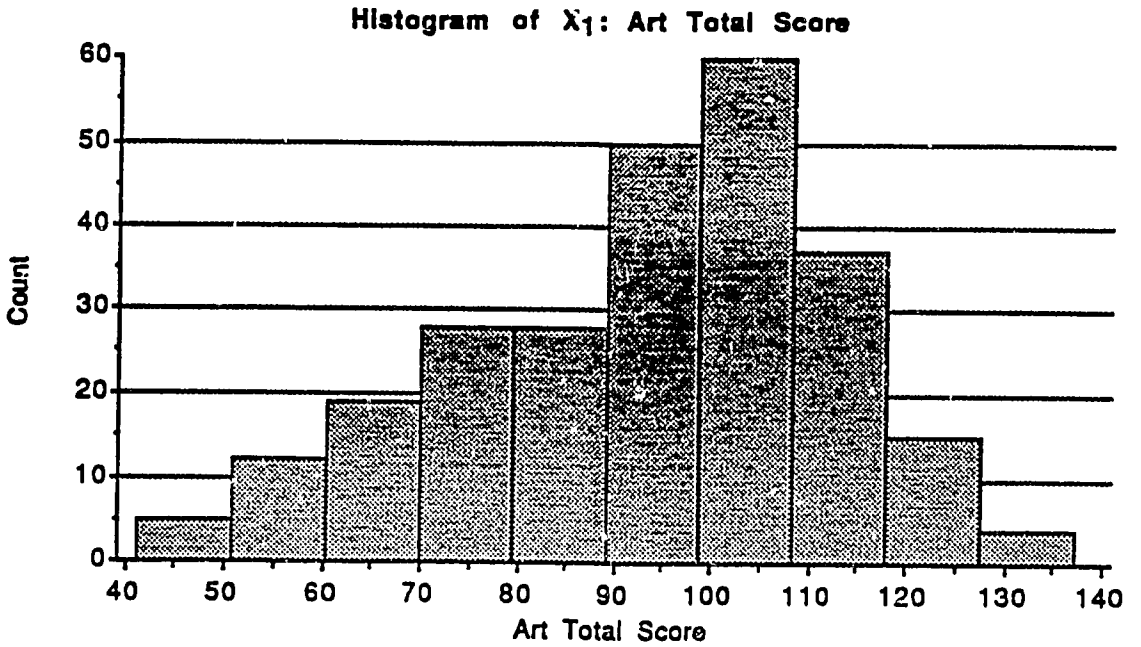

Figure 3. 1987 Two-Year Art JCE Total Score Distribution: Mean = 92.9

Score Statistics. Table 1 shows the means, standard deviations, and intercorrelations among the 1987 Art JCE scores. Two features of the scores stand out. There is a low correlation of 0.25 between Paper 1 and Portfolio scores, and the relative standard deviations of scores for Paper 1 (5.6) and the Portfolio (16.8) indicate that the Portfolio score considerably outweighs the Paper 1 score in determining the Total score. The effective or "real" weights for Paper 1 and Portfolio scores are in the ratio of 23:127 out of 150 marks (5.5 to 1) rather than 40:110 (2.75 to 1) which one might erroneously infer from their relative maximum marks.[1] The dominant weighting of the Portfolio score in the Total score is also reflected in the high correlation of 0.96 between the Portfolio score and the summed Total score.

Table 1

1987 Two-Year JCE Score Statistics

Score	Mean	S.D	Intercorrelations (1)	(2)	(3)
(1) Paper 1	28.1	5.6	1.00	--	--
(2) Portfolio	64.8	16.8	0.25	1.00	--
(3) Total	92.9	19.1	0.51	0.96	1.00

[1] The ratio 23:127 is equal to the relative covariances of Paper 1 and Portfolio scores with the Total score, which sum to the variance of the Total score.

Item Statistics for Paper 1

Table 2 shows two statistics for each of the 40 items in Paper 1. These are p and r.

p = the proportion of examinees who chose the correct answer for each item;
r = the correlation of each item's score with the total number correct score for Paper 1.

Average p-value is 0.70; average r-value is 0.34. The most difficult item in Paper 1 is Item 36 with the lowest p-value of 0.32. The easiest item is Item 3 with a p-value of 0.98. Item 36 (the most difficult with p = 0.32) presents four illustrations of African art pieces and asks examinees to identify which one represents abstract art. Item 3 (the easiest with p = 0.98) asks examinees where colour is used in everyday life as decoration -- the four multiple-choice options are "compound walls, a Kgotla chair, a cattle kraal," and a "borehole."

The most discriminating item (the item that correlates highest with overall score on Paper 1) is Item 5 with an r-value of 0.50. Item 5 asks for the correct definition of the concept "perspective"; options are "showing depth on a flat surface, the emphasis of one part over another, repetition of elements," and "intensity of colour." Item 2 is least discriminating (students who do well or poorly overall answer the item equally correctly) with an r-value of 0.09. Item 2 asks whether lines express "depth, movement, contour," or "all of the above."

Table 2

1987 Two-Year Art JCE Item Statistics for Paper 1

Item	p	r	Item	p	r
1	.87	.38	21	.87	.28
2	.58	.09	22	.76	.38
3	.98	.20	23	.50	.38
4	.76	.32	24	.60	.42
5	.69	.50	25	.85	.28
6	.66	.40	26	.64	.18
7	.46	.40	27	.41	.27
8	.64	.27	28	.47	.32
9	.95	.25	29	.75	.41
10	.80	.42	30	.61	.38
11	.85	.45	31	.67	.25
12	.70	.34	32	.85	.39
13	.93	.43	33	.75	.24
14	.61	.44	34	.75	.38
15	.80	.26	35	.83	.28
16	.89	.44	36	.32	.19
17	.61	.37	37	.55	.40
18	.60	.30	38	.83	.28
19	.79	.48	39	.73	.31
20	.81	.37	40	.46	.27

Figure 4 shows a scatterplot of Item p-values and Item r-values which are listed in Table 2. Overall, 14 of 40 items were relatively easy (p-values of 0.80 or more); 26 items were moderately difficulty (p-values between 0.30 and 0.80). Five items showed low discrimination (r-values below 0.25). Whether item statistics are desirable or not is a matter for JCE art test constructors to decide. If Paper 1 is viewed as a competency test, then desirable item statistics would show high p-values; low r-values are acceptable as long as they are positive. If Paper 1 is intended to reliably discriminate among students, then desirable item statistics would show moderately difficult p-values and r-values above 0.25.

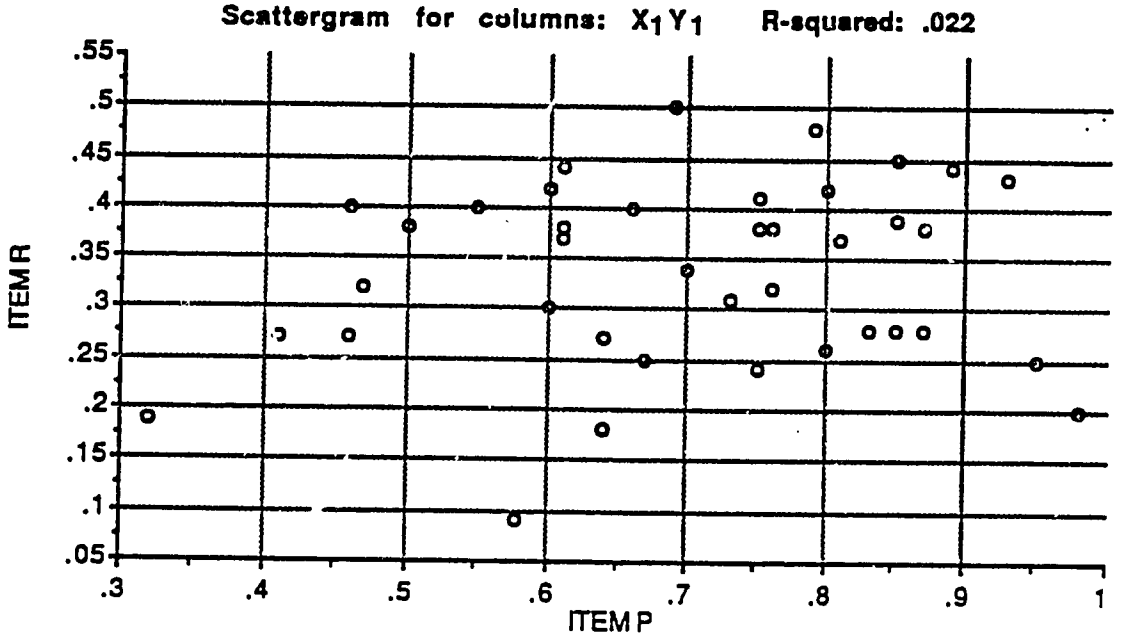

Figure 4. Scatterplot of Item p and Item r for Paper 1 (taken from Table 2).

School Averages

Average scores for Paper 1, the Portfolio, and the Total Score are shown in Table 3 for students in each of the seven schools where art was introduced in 1986. Mean differences among the schools were highly significant; analyses of variance F-tests of equal school means were F = 38.0 for Paper 1, F = 22.0 for the Portfolio, and F = 26.7 for the Total score. All F-tests were significant at p = .0001.

Table 3

School Mean Scores for 1987 Art JCE

School	Paper 1	Portfolio	Total
A	29.5	61.7	91.2
B	30.8	79.7	110.5
C	32.9	76.1	109.0
D	29.1	72.8	101.9
E	20.8	73.9	94.7
F	25.2	51.0	76.2
G	32.0	70.3	102.3
F-test	38.0	22.0	26.7
p	.0001	.0001	.0001

Summary and Discussion

Conclusions regarding the results of the first administration of the Two-Year Art JCE are:

- Distributions of scores for Paper 1, the Portfolio, and the Total score spread well across their respective ranges but showed negative skew, which means that scores were moderately high. In terms of maximum score, respective means for the three scores were 70 per cent, 60 per cent, and 62 per cent.

- The Portfolio score outweighs the score for Paper 1 by the ratio of 127:23 (5.5 to 1) when they are summed to derive the Total score. It is wrong to think that their relative weights are in the ratio of 110:40 (2.75 to 1) reflected by their respective maximum scores. An intended weighting ratio can be specified and can be precisely controlled by first standardizing and then weighting their standard scores before they are summed. Otherwise, relative weighting is uncertain and determined by whatever statistical characteristics are shown by the obtained score distributions. This is true for any summed total score in any subject.

- Intended relative weighting for the two scores is a decision best left to JCE Art test constructors. This decision should be based upon the purposes to be served by Paper 1 and the Portfolio in evaluating student achievement on the Art JCE.

- One can consider two possible ways in which Paper 1 and the Portfolio may be regarded.

- The evaluation purposes of Paper 1 and the Portfolio may be regarded as similar – to discriminate among students both in terms of "knowing about art" measured by Paper 1 and "knowing how to do art" shown by artwork pieces submitted in the Portfolio. Moderate p-values and r-values above 0.25 would then be desirable for Paper 1, and Portfolio scores should be spread out.

- On the other hand, Paper 1 and the Portfolio may be regarded as serving distinct evaluation purposes. The multiple-choice Paper 1 could be regarded as a knowledge competency test – a "knowing about art" test that covers basic elements and principles that all teachers are expected to teach and all students are expected to learn. Item statistics for Paper 1 would show high p-values; the score distribution for Paper 1 would show negative skew with individual scores piling up at the high end of the score scale. The purpose of the Portfolio would be to discriminate among students in terms of those who know "how to do art."

One final recommendation is that the Art JCE Portfolio raters continue to pay close attention to setting clear and observable criteria for rating submitted artwork and to establishing acceptably high inter-rater reliability. Training sessions for Portfolio markers should be carried out.

A Preliminary Analysis: The 1987 Administration of the IEA Tests to Two-Year and Three-Year JC Completers

Research Note 88-05

July 1988

Research and Testing Centre
Curriculum Development and Evaluation
Ministry of Education
Government of Botswana

A Preliminary Analysis The 1987 Administration of the IEA Tests to Two-Year and Three-Year JC Completers

Background

As part of the Evaluation Task Force plans to get some initial feedback on the performance of the New Two-Year Junior Secondary Programme , a battery of four IEA tests, developed by the International Association for the Evaluation of Educational Achievement, was administered in 52 Secondary schools in May ,1987. The tests were administered to the Two-Year and Three-Year Junior Certificate completers in each school. The tests in the battery were Mathematics, Reading Comprehension, Word Knowledge and Science. These were the same tests used in the 1976 Primary and Secondary Education Survey for the National Commission on Education.

The administration of the tests formed part of the teaching practice exercise for the 1987 third-year students at the Molepolole College of Education Students sent to each of the 52 schools were therefore responsible for the administration of the tests. This report presents a preliminary analysis of the IEA test results. Future reports will examine the relationships at the school mean level of achievement on the IEA tests and JC Examinations.

Testing was carried out in 52 secondary schools. However, the numbers of senior or junior secondary schools where testing was conducted varied, depending on the placement of third year students. Because of local contingencies, all four tests were not administered in all schools. IEA Reading Comprehension data were missing for most schools with Three-Year JCE completers, so this test is not analysed. Table 1 shows the numbers of schools in which each test was administered to Two-Year and Three-Year JCE completers.

Table 1

1987 IEA Test Administration By Type of School And By Two-Year Or Three-Year JCE Completion

IEA Test	Junior Secondary		Senior Secondary	
	2-Year	3-Year	2-Year	3-Year
Word Knowledge	31	29	17	15
Science	35	31	17	17
Mathematics	34	29	16	15

IEA Tests

Word Knowledge. The Word Knowledge test consists of 40 items each of which presents a pair of words. Examinees indicate whether each word pair is the same or opposite in meaning. Score is number correct.

Science. This is a 34 item multiple choice test measuring knowledge, concepts and applications in Science. Each item presents five options. The total score for each item is the number of items answered correctly.

Mathematics. The Mathematics test consists of 24 multiple choice items on basic Arithmetic and Mathematics skills . Each item presents five options. The total score for each student is the number of items answered correctly.

Procedures

School mean performance is the unit of analysis for each test. Data are analysed in three ways:

(1) A two-way analysis of variance was carried out to determine whether school means significantly differed according to School Type (Junior or Senior Secondary), according to Year of JCE Completion (Two-Year or Three-Year), and to determine whether there were Type by Year effects. Thus the analysis examines the two main effects for TYPE AND YEAR, as well as the TYPE X YEAR interaction effect.

(2) Scheffé tests were carried out to test for significant mean differences between all of the six possible TYPE x YEAR pairwise comparisons.

(3) School means on each test were classified as being above or below the median of the combined distribution for junior and secondary schools. A chi-square analysis was then run to determine whether there were significant differences in the proportions of school means above the median or below the median in each Type X Year category.

ANOVA Results

IEA Word Knowledge. Table. 2 shows an analysis of variance summary for Word Knowledge. There is no YEAR x TYPE of SCHOOL interaction effect. The mean of the Junior Secondary Schools (23.8) was significantly different from the mean of the Senior Secondary Schools (24.7). This mean difference is approximately 0.7 standard deviations (the estimated within-group standard deviation is the square root of the mean square for error which is 1.74). There is also a significant difference between the mean of the 2-year Junior Certificate completers (23.6) and the mean of the 3-year Junior Certificate completers (24.7), a mean difference of 0.8 standard deviations.

Table 2

Analysis of Variance Testing for School and Type Main Effects and for School by Type Interaction for IEA Word Knowledge

Effect	df	Mean Square	F-Test	P-value
TYPE of SCHOOL	1	18.09	10.41	.0018
YEAR of JCE	1	23.81	13.69	.0004
TYPE X YEAR	1	.38	.22	.6427
Error	88	1.74		

Scheffé tests of the significance of the mean differences for all possible pairs of the four groups were calculated. The Scheffé tests of all possible two-group comparisons allows for a more precise interpretation of the differences detected by the analysis of variance. The results are shown in Table 3, JSS and SSS stand for Junior and Senior Secondary Schools, respectively, and the means for each of the four groups are listed in the last column. The 2-Year Junior Secondary School mean (23.4) was found to be significantly different from the 3-Year Senior Secondary School mean (25.4).

Table 3

Scheffé Tests of Pairwise Comparisons on IEA Word Knowledge

Group	JSS-2	JSS-3	SSS-2	SSS-3	Mean
JSS-2		---	---	*	23.4
JSS-3			---	---	24.3
SSS-2					24.2
SSS-3					25.4

IEA Science. Table 4 shows the analysis of variance results for IEA Science. There is no significant TYPE of SCHOOL x YEAR interaction effect. There were significant mean differences between the YEAR groups: the IEA Science mean for the 2-year JC completers was 11.7, compared to 12.7 for the 3-year JC completers. There were also significant mean differences between the TYPE groups: the IEA Science mean for the JSS JC completers was 11.8, compared to 12.8 for the SSS JC completers.

Table 4

Analysis of Variance Testing for School and Type Main Effects and for School by Type Interaction for IEA Science

Effect	df	Mean Square	F-Test	P-value
TYPE of SCHOOL	1	18.45	22.95	.0001
YEAR of JCE	1	25.18	31.31	.0001
TYPE X YEAR	1	2.75	3.42	.0676
Error	96	0.80		

There were four significant mean differences shown by Scheffé tests reported in Table 5. These were between the 2-year and 3-year JC completers in junior secondary schools, and between the 3-year JC completers in senior secondary school and each of the other three groups.

Table 5

Scheffé Tests of Pairwise Comparisons on IEA Science

Group	JSS-2	JSS-3	SSS-2	SSS-3	Mean
JSS-2		---	*	*	11.5
JSS-3			---	*	12.2
SSS-2				*	12.1
SSS-3					13.5

IEA Mathematics. The results for the analysis of variance for IEA Mathematics are shown in Table 6. There is a significant TYPE of SCHOOL x Year interaction effect. There is also a significant TYPE difference between the Junior Secondary School mean (11.0) and the Senior Secondary School mean (12.0), and between the means for 2-Year completers (11.0) and 3-Year completers mean (11.8). The school TYPE differences (JSS vs SSS) are approximately 0.9 standard deviations and the YEAR differences (2-Year vs 3-Year) are approximately 0.7 standard deviations.

Table 6

Analysis of Variance Testing for School and Type Main Effects and for School by Type Interaction for IEA Mathematics

Effect	df	Mean Square	F-Test	P-value
TYPE of SCHOOL	1	18.84	14.64	.0002
YEAR of JCE	1	21.38	16.61	.0001
TYPE X YEAR	1	8.20	6.37	.0134
Error	90	1.29		

The two- group Scheffé comparisons in Table 7 show that the mean for the Senior Secondary 3-year JC completers was significantly different from all other group means. The 3-year JC completers in Senior Secondary Schools performed significantly better(mean=12.8) than did the 2-year JC completers in Senior Secondary Schools (mean=11.2), the 3-Year JC completers in Junior Secondary Schools (mean=11.3), and the 2-Year JC completers in Junior Secondary Schools (mean=10.9).

Table 7

Scheffé Tests of Pairwise Comparisons on IEA Mathematics

Group	JSS-2	JSS-3	SSS-2	SSS-3	Mean
JSS-2		---	---	*	10.9
JSS-3			---	*	11.3
SSS-2				*	11.2
SSS-3					12.8

CHI-SQUARE ANALYSIS

School means for each IEA test for the four groups of examinees were indexed as either above or below the median of the distribution of all school means for that test. Tables 8 through 10 show the classification of schools above or below the median for each IEA test.

IEA Word Knowledge. Table 8 shows the percentage of school means for each of the four groups that are below or above the overall median for IEA Word Knowledge.

Table 8

IEA Word Knowledge School Means Above or Below Overall Median Classified by Year and Type

1987	Two-Year JC				Three-Year JC			
School	Junior		Senior		Junior		Senior	
Means	N	%	N	%	N	%	N	%
Above Median	9	29 %	7	41 %	17	59 %	13	87 %
Below Median	22	71	10	59	12	41	2	13
All Means	31	100 %	17	100 %	29	100 %	15	100 %

Chi-Square (3 df) = 14.91; $p < 0.01$

When all groups are ranked according to the proportion of school means above the overall IEA Word Knowledge median, the 3-Year JC completers in Senior Secondary Schools and in Junior Secondary Schools had higher proportions than the both 2-Year groups in Senior and Junior Schools.

3-Year	SSS	87%
3-Year	JSS	59%
2-Year	SSS	41%
2-Year	JSS	29%

Comparisons between school TYPES show that the percentage of Senior Secondary Schools above the median exceeded the percentage of Junior Secondary Schools within both the 2-Year groups and the 3-Year groups. Within the 2-Year group, 41% of the Senior Secondary means were above the overall median compared with 29% of the Junior Secondary means. In the 3-Year group, the comparative figures were 87% of the Senior Secondary Schools and 59% of the Junior Secondary Schools.

The 2-Year vs 3-Year comparisons show that in the Junior Secondary Schools, the 3-Year group had a higher proportion (59%) of school means above the overall median than the 2-Year group (29%). In the Senior Secondary Schools, the 3-Year group had 87% of the school means above the overall median with 41% of the school means falling above the median for the 2-year group.

IEA Science. Table 9 shows the proportions of school means below and above the overall median for IEA Science. There were significant differences in the category percentages (chi-square = 20.01; $p < .01$).

Table 9

IEA Science School Means Above or Below Overall Median Classified by Year and Type

1987 School Means	Two-Year JC				Three-Year JC			
	Junior		Senior		Junior		Senior	
	N	%	N	%	N	%	N	%
Above Median	10	29 %	8	47 %	17	55 %	16	94 %
Below Median	25	71	9	53	14	45	1	16
All Means	35	100 %	17	100 %	31	100 %	17	100 %

Chi-Square (3 df) = 20.01; $p < 0.01$

When all groups are ranked according to the proportion of means above the overall IEA Science median, the 3-Year JC completers in Senior Secondary Schools and in Junior Secondary Schools had higher proportions than both 2-Year groups in Senior and Junior Schools.

3-Year	SSS	94%
3-Year	JSS	55%
2-Year	SSS	47%
2-Year	JSS	29%

Comparisons between school TYPES shows that the Sennior Secndary Schools performed better than Junnior Secondary Schools within both the 2-Year groups and the 3-Year groups. Within the 2-Year group 47% of the Sennior Schools means were above the overall median with 29% of the Junnior Schools means above the overall median. In the 3-Year group, 94% of the Sennior schools were above the averall median with 55% of the Junnior schools means above the overall median.

The 2-Year vs 3-Year comparisons show that in the Junior Secondary Schools the 3-Year group had a higher proportion (55%) than the 2-Year group (29%). In the Senior Secondary Schools, the 3-Year group had 94% of the school means above the overall median with 47% of the school means falling in the same category for the 2-year group.

IEA Mathematics. Table 10 shows the proportions of school means below and above the overall median for IEA Mathematics . There were significant differences in the category percentages (chi-square = 22.81; p < .01).

Table 10

IEA Mathematics: School Means Above or Below Overall Median Classified by Year and Type

1987 School Means	Two-Year JC Junior N	%	Two-Year JC Senior N	%	Three-Year JC Junior N	%	Three-Year JC Senior N	%
Above Median	9	26 %	9	56 %	14	48 %	15	100 %
Below Median	25	74	7	44	15	52	--	--
All Means	34	100 %	16	100 %	29	100 %	15	100 %

Chi-Square (3 df) = 22.81; p < 0.01

The ranking of the proportion of means above the overall IEA Mathematics median for the four groups shows that the Senior Secondary Schools means (2-Year and 3-Year) had higher proportions above the median than Junior Secondary Schools means (2-Year and 3-Year).

3-Year	SSS	100%
2-Year	SSS	56%
3-Year	JSS	48%
2-Year	JSS	26%

Comparisons between school TYPES shows that the Senior Secondary Schools showed a higher proportion above the median than Junior Secondary Schools within both 2-Year and 3-Year groups. Within the 2-Year group 56% of the Senior Schools means were above the overall median and 26% of the Junior Schools. In the 3-Year group, 100% of the Sennior schools were above the averall median with 48% of the Junior schools means above the overall median.

The 2-Year vs 3-Year comparisons show that in the Junior Secondary Schools the 3-Year group had a higher proportion (48%) of school means above the overall median than the 2-Year group (26%). In Senior Secondary Schools, the 3-Year group had 100% of the school means above the overall median with 56% of the school means falling in the same category for the 2-Year group.

SUMMARY

This school-level analysis of the 1987 IEA test data for 2-Year and 3-Year JC completers in Senior and in Junior Secondary schools supports the following conclusions.

(1) On all IEA tests, the ANOVA analyses showed significant main effects for YEAR (3-Year vs. 2-Year completers) and TYPE (Senior vs. Junior Secondary Schools).

(2) Seven of the eight Scheffé pairwise comparisons showing significant mean differences favoured the 3-year Senior Secondary group over other YEAR x TYPE groups. The 3-year Secondary School means on IEA Science and IEA Mathematics were significantly higher than the means of the 2-Year Senior Secondary groups and both the 2-Year and 3-Year Junior Secondary groups. The 3-Year Secondary School mean on IEA Word Knowledge was significantly higher than the mean for the 2-Year Junior Secondary group.

(3) The one other significant Scheffé comparison showed a mean difference favouring the 2-Year Senior Secondary group over the 2-Year Junior Secondary group on IEA Science.

(4) The chi-square analysis compared TYPE x YEAR groups in terms of the proportion of TYPE x YEAR school means above or below the median of the overall distribution for each IEA test. Essentially the same findings were revealed. For IEA Sciencce and IEA Word Knowledge, a greater proportion of the 3-Year Junior and Senior means fell above the median than did 2-Year Junior and Senior means. On IEA Mathemetics, a greater proportion of the Senior Secondary 2-Year and 3-Year means fell above the median than did Junior Secondary 2-Year and 3-Year means.

Following this preliminary analysis will be a school level comparison of performances on the IEA tests and the Junior Certificate Examinatons.

1987 Two-Year JC Multiple-Choice Item and Total Score Statistics

Research Note 88-06

July 1988

Research and Testing Centre
Curriculum Development and Evaluation
Ministry of Education
Government of Botswana

1987 Two-Year JC Multiple-Choice Item and Total Score Statistics

Previous RTC Research Notes have pointed out the relationships between test item and test score statistics. For JC Examinations, Paper 1 consists of n multiple-choice items each scored 1 if answered correctly and 0 otherwise; the total test score is the sum of the n item scores.

Customary item statistics p_i (the proportion of examinees answering the i-th item correctly) and r_{it} (the correlation between the i-th item score and the total test score) are related to the total test score mean and standard deviation as shown below.

$$\Sigma p_i = \text{the total test score mean;} \qquad (1)$$

$$\Sigma r_{it} [p_i (1-p_i)]^{1/2} = \text{the total test score standard deviation.} \qquad (2)$$

This relationship is of more than algebraic interest. It implies that test constructors who build tests with known item p_i and r_{it} values are able to control the mean and standard deviation of the total score on tests they construct. Together with control of test content, this means that equivalent tests in a subject can be constructed from one year to the next, that the performance of examinees can be compared in any single subject from year to year, and that, if desired, multiple choice tests across subjects in any one year can be made statistically equivalent.

Control is made possible by selecting items that are indexed in files according to their content and statistical properties. This is of special importance during the rapid expansion of secondary schools; many students are provided educational opportunity who would not otherwise continue schooling past the primary level. The construction of equivalent tests would provide the Ministry of Education with the means to establish baseline JC achievement scales and to monitor future changes in JC achievement results.

Reported Data

This note examines the Paper 1 multiple choice tests of the Two-Year JC Examinations that were administered in late 1987. Item and test score statistics are presented. These reveal variation in the statistical properties of multiple choice tests in different subjects. Test constructors in each subject may examine the data to decide whether the test distributions are those they desire, or if not, how steps may be taken to change distributions in a desired direction at the time that items are written or selected for future JC tests.

In reporting the item statistics, item p_i and item r_{it} values are cross-classified. Item r_{it} values are classified into two categories; item r_{it} values above 0.25 are customarily considered "good."

r_{it} less than 0.25	= Poor discrimination
r_{it} 0.25 and above	= Good discrimination

Item p_i values are classified into five categories; item p_i values between 0.20 and 0.80 are customarily considered "good."

Less than 0.20	= Very difficult
0.20-0.39	= Moderately difficult
0.40-0.59	= Average
0.60-0.79	= Moderately easy
0.80 and above	= Very easy

The important point is to select items that produce test distributions that serve intended purposes. Content coverage is a paramount consideration. However, one assumes at present that multiple choice JC Examinations are used for normative measurement -- to make reliable discrimination among the candidates' levels of achievement. One would prefer flat test distributions so the test standard deviation should not be small. For this purpose, item p_i values should be spread but should average around 0.60 to guard against obtained low scores that result from guessing. Item r_{it} values will show moderate correlation levels since items span a wide range of content, but selecting items with too high a correlation level will narrow content coverage. Low item r_{it} values indicate poor discrimination and reduce test reliability. It is for these reasons that item p_i values between 0.20 and 0.80 and item r_{it} values above 0.25 are considered "good." If in future, JC multiple choice tests should become more competency-based, then different rules for item selection would operate.

JC Examination Paper 1 Item and Total Test Score Statistics

Table 1 lists item p_1 and item r_{it} values for the Two-Year JC Paper 1 multiple choice tests in English, Setswana (also with a multiple choice test for Paper 2), Mathematics, Science, Social Studies, Home Economics, Religious Education, and Agriculture; also shown are total score means and standard deviations. Technical Studies has two different multiple choice tests, but the data are not reported since the computer printouts combined statistics for both. These tests will be reported in future.

Table 1

Item and Total Test Score Statistics for 1987
JCE Paper 1Multiple Choice Tests

JCE Subject Test Multiple Choice	Item p_1 Mean	Item r_{it} Mean	Test Score Mean	S.D.
English (40 items)	0.64	0.27	25.6	4.5
Setswana 1 (40 items)	0.60	0.26	22.8	4.2
Setswana 2 (40 items)	0.63	0.28	25.1	4.6
Mathematics (40 items)	0.33	0.27	13.4	4.9
Science (60 Items)	0.57	0.26	34.5	7.2
Social Studies (50 items)	0.65	0.28	32.8	6.1
Home Economics (40 items)	0.53	0.25	20.6	4.4
Religious Education (60 items)	0.61	0.32	36.9	8.4
Agriculture (50 items)	0.50	0.25	25.1	5.8

Average item p_i values range from 0.33 for Mathematics to 0.65 for Social Studies; average item r_{it} values range from 0.25 for Home Economics and for Agriculture to 0.32 for Religious Education. Tests presently, of course, do not show all items with what are customarily considered "good" item statistics -- that is, those with item p_i values between 0.2 and 0.8 and item r_{it} values greater than 0.25. However, should files containing items with known statistical and content characteristics be available at the time of test construction, items with these "good" statistical characteristics could be selected, yielding tests with desired means and standard deviations.

English Paper 1

Figure 1 below shows the distribution of scores for JC Paper 1 in English. Table 2 shows a cross-classification of item p_i values and item r_{it} values for this test.

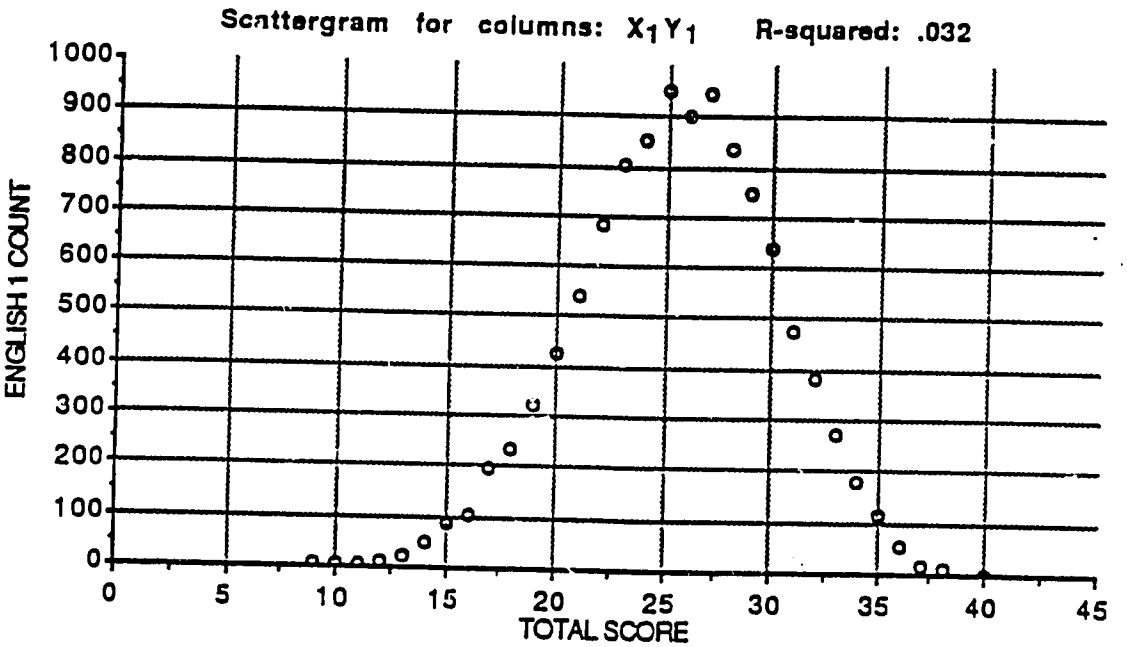

Fig. 1. Distribution of 1987 Two Year JC Paper 1 in English. (Mean=25.6; S.D.=4.5)

Table 2

Cross-Classification of Item p_1 and Item r_{it} Values for 1987 JC Paper 1 in English

Item p_i	N	Item r_{it} values N<0.25	N≥0.25
Less than 0.20	2	1	1
0.20-0.39	4	0	4
0.40-0.59	11	5	6
0.60-0.79	10	2	8
0.80 and above	13	8	5
Totals	40	16 40%	24 60%

The English multiple choice test was somewhat easy, with nearly one-third of its item p_i values at 0.80 or above. Sixteen items (40 per cent of the total) showed low discrimination with item r_{it} values below 0.25. Of these, 8 items were also very easy with p_i values of 0.80 or above. These should be revised if the objectives they measure are tested in future. The test score distribution shows only a slight negative skew. Revising or replacing the very easy items with low discrimination would lower the mean somewhat and flatten the distribution. This would result in higher test reliability.

Setswana Papers 1 and 2

Figure 2 below shows the distribution of scores for JC Paper 1 in Setswana. Table 3 shows a cross-classification of item p_i-values and item r_{it} values for this test.

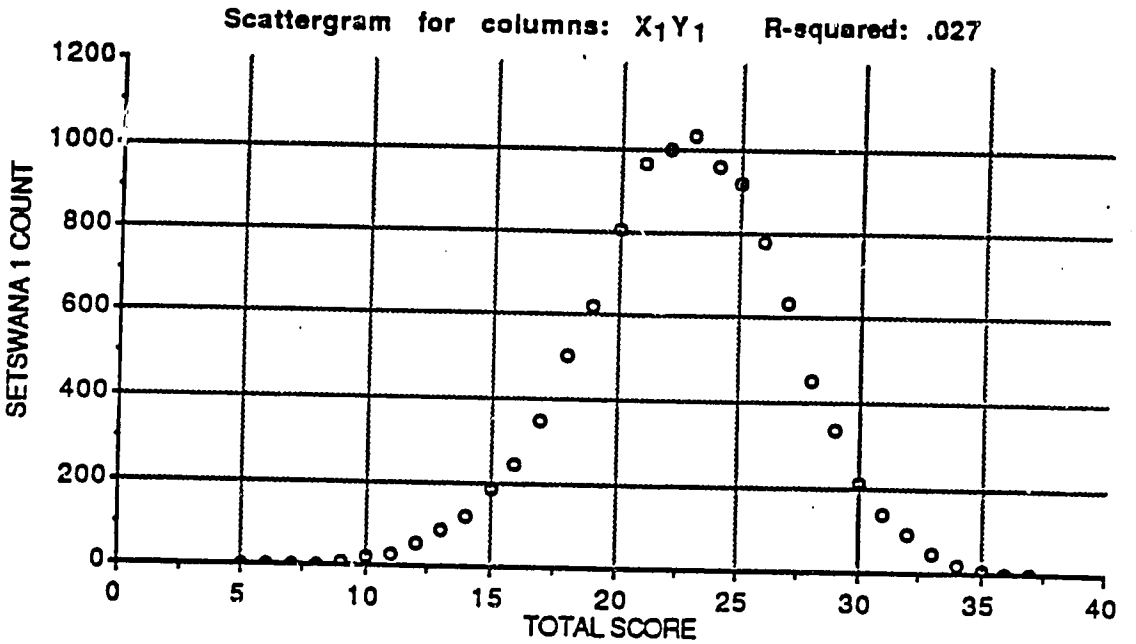

Fig. 2. Distribution of 1987 Two Year JC Paper 1 in Setswana 1. (Mean=22.8; S.D.=4.2)

Tab 3

Cross-Classification of Item p_1 and Item r_{it} Values for 1987 JC Paper 1 in Setswana

Item p_i	N	Item r_{it} values N<0.25	N≥0.25
Less than 0.20	1	1	0
0.20-0.39	9	5	4
0.40-0.59	8	3	5
0.60-0.79	12	1	11
0.80 and above	10	7	3
Totals	40	17 42%	23 58%

The Setswana Paper 1 multiple choice test was a bit more difficult than the English Paper 1, with one-fourth of its p_i values at 0.80 or above. Of these 10 items, 7 had item r_{it} values below 0.25. Nine of the 29 items with item p_i values between 0.20 and 0.79 also showed low discrimination with item r_{it} values below 0.25, and should be examined for revision. The test could be given a higher ceiling through revising these items.

Figure 3 below shows the distribution of scores for JC Paper 1 in Setsana. Table 4 shows a cross-classification of item p_i-values and item r_{it} values for this test.

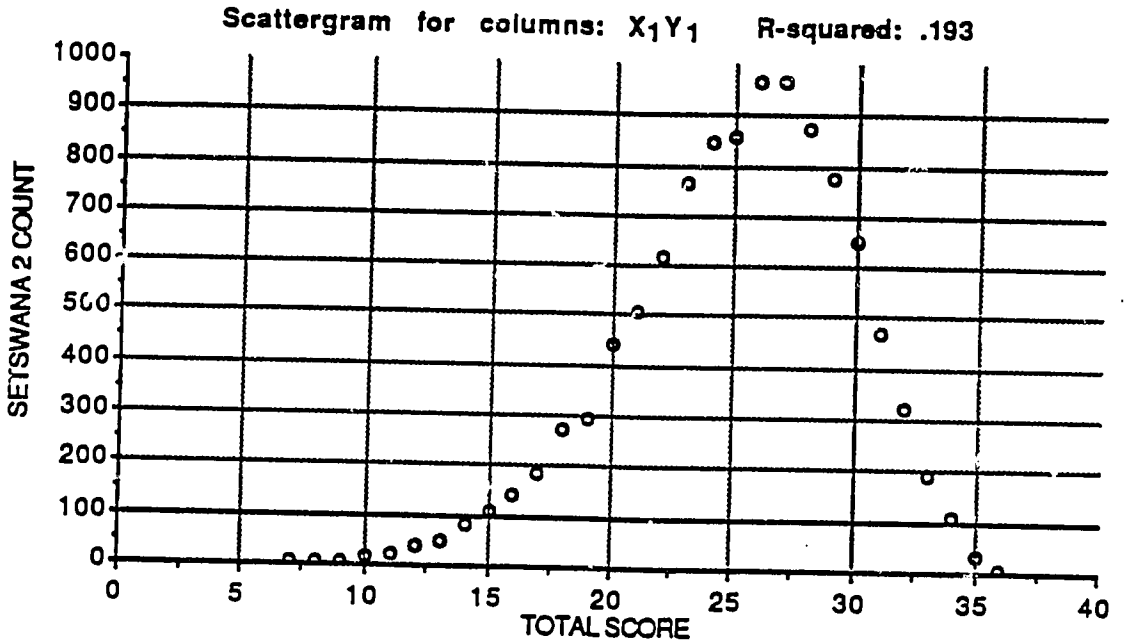

Fig. 3. Distribution of 1987 Two Year JC Paper 2 in Setswana. (Mean=25.1; S.D.=4.6)

Table 4

Cross-Classification of Item p_1 and Item r_{it} Values for 1987 JC Paper 2 in Setswana

Item p_i	N	Item r_{it} values N<0.25	N≥0.25
Less than 0.20	0	0	0
0.20-0.39	7	5	2
0.40-0.59	10	5	5
0.60-0.79	11	1	10
0.80 and above	12	1	11
Totals	40	12 30%	28 70%

The Setswana Paper 2 multiple choice test is slightly easier than Setswana Paper 1. Its items also show somewhat better discrimination. Twelve items showed item p_i values of 0.80 or above, all but one showed good discrimination with item r_{it} values above 0.25. Eleven of the 28 items with p_i values between 0.20 and 0.79 showed item r_{it} values below 0.25 and should be examined for revision. The test score distribution is negatively skewed, but replacement of the very easy items with more difficult ones would tend to lower the mean, flatten the distribution, and enable the test to discriminate better in the middle range of ability.

Mathematics Paper 1

Figure 4 below shows the distribution of scores for JC Paper 1 in Mathematics. Table 5 shows a cross-classification of item p_i-values and item r_{it} values for this test.

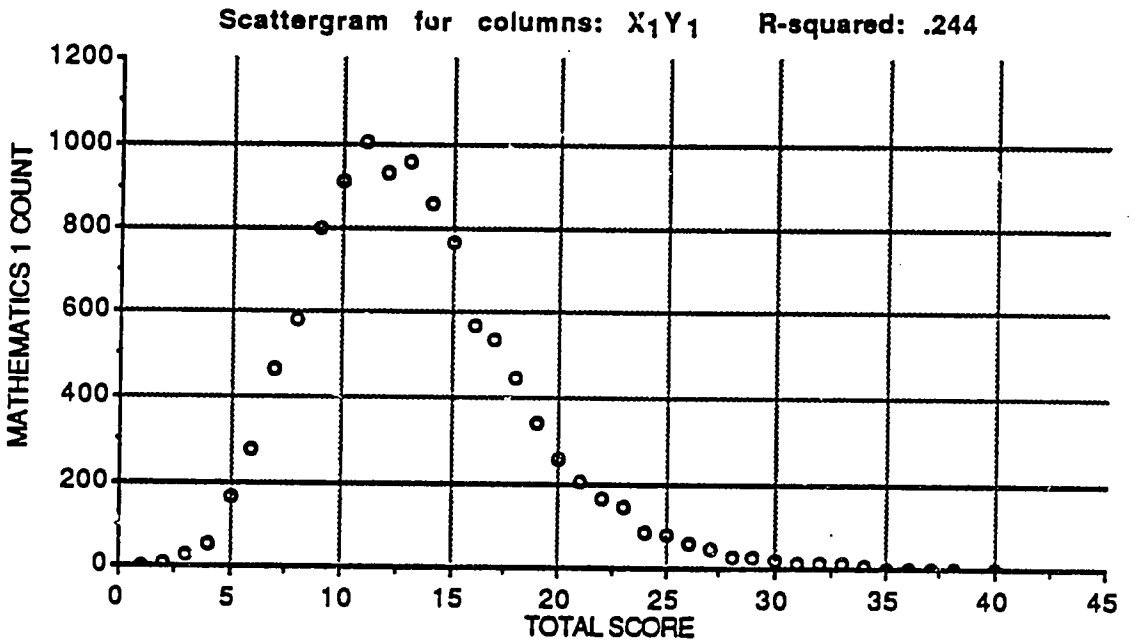

Fig. 4. Distribution of 1987 Two Year JC Paper 1 in Mathematics. (Mean=13.4; S.D.=4.9)

Table 5

Cross-Classification of Item p_i and Item r_{it} Values for 1987 JC Paper 1 in Mathematics

Item p_i	N	Item r_{it} values N<0.25	N≥0.25
Less than 0.20	9	6	3
0.20-0.39	16	5	11
0.40-0.59	13	2	11
0.60-0.79	2	0	2
0.80 and above	0	0	0
Totals	40	13 32%	27 68%

The Mathematics Paper 1 multiple choice test is the most difficult of all those considered. Only two items have moderately easy p_i values of 0.60 or higher, with none above 0.80. Of the nine items with p_i values below 0.20, six showed poor discrimination with item r_{it} values below 0.25. Seven of the 16 items with item p_i values between 0.20 and 0.59 have item r_{it} values below 0.25. The high negative skew of the test score distribution reflects best discrimination among higher achieving candidates. Easing the difficulty level of the test would increase its discrimination among the average and lower achieving candidates, and would raise the test mean above the range of scores where guessing may operate.

Science Paper 1

Figure 5 below shows the distribution of scores for JC Paper 1 in Science. Table 6 shows a cross-classification of item p_i-values and item r_{it} values for this test.

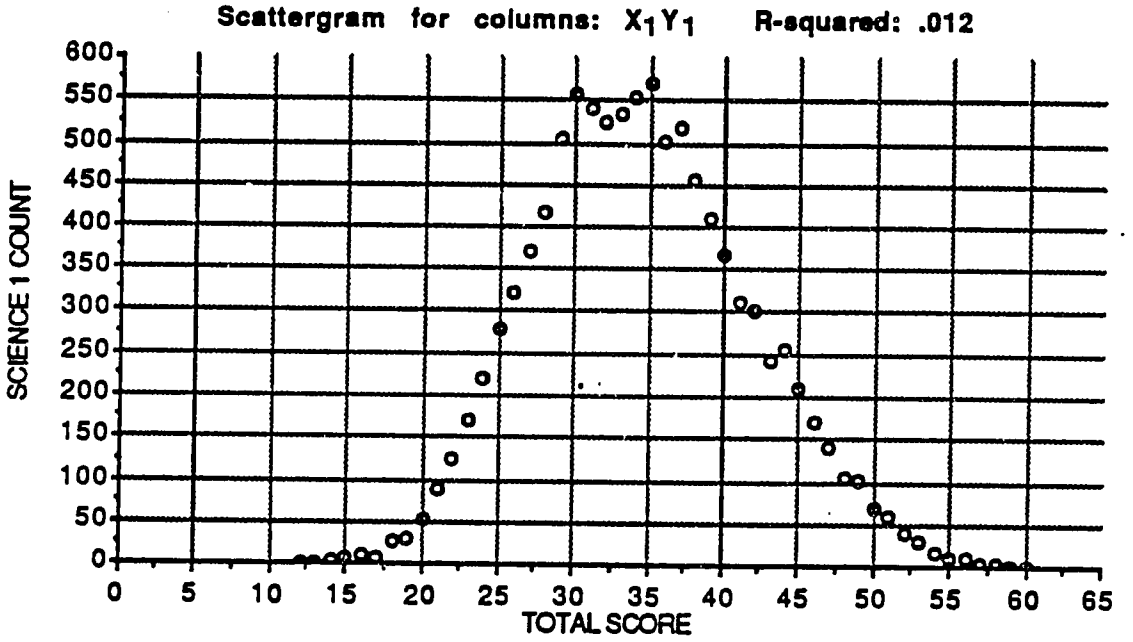

Fig. 5 Distribution of 1987 Two Year JC Paper 1 in Science. (Mean=34.5; S.D.=7.2)

Table 6

Cross-Classification of Item p_1 and Item r_{it}
Values for 1987 JC Paper 1 in Science

Item p_i	N	Item r_{it} values N<0.25	N≥0.25
Less than 0.20	1	1	0
0.20-0.39	13	7	6
0.40-0.59	19	6	13
0.60-0.79	16	4	12
0.80 and above	11	7	4
Totals	60	25 42%	35 58%

The Science test shows a slight positive skew reflecting the effect of 11 very easy items with p_i-values of 0.80 and above. Of these, seven also showed low discrimination with r_{it}-values below 0.25. These should be examined for possible revision. Making the test somewhat more difficult and replacing items with low discrimination would improve its reliability and discrimination over a broader ability range.

Social Studies Paper 1

Figure 6 below shows the distribution of scores for JC Paper 1 in Mathematics. Table 7 shows a cross-classification of item p_i-values and item r_{it} values for this test.

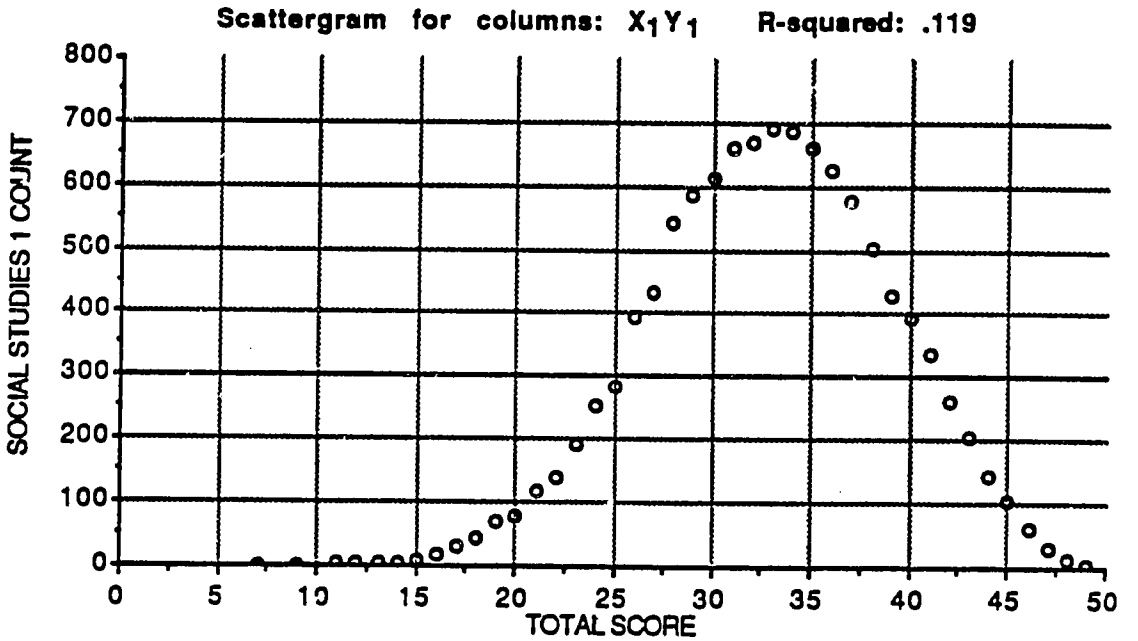

Fig. 6 Distribution of 1987 Two Year JC Paper 1 in Social Studies. (Mean=32.8; S.D.=6.1)

Table 7

Cross-Classification of Item p_1 and Item r_{it}
Values for 1987 JC Paper 1 in Social Studies

		item r_{it} values			
Item p_i	N	N<0.25		N≥0.25	
Less than 0.20	0	0		0	
0.20-0.39	7	4		3	
0.40-0.59	13	2		11	
0.60-0.79	16	3		13	
0.80 and above	14	8		6	
Totals	50	17	34%	33	66%

This is a relatively easy test with a slight negative skew. There were no items with p_i-values below 0.20, and 7 items with p_i-values between 0.20 and 0.39. Seventeen of the Social Studies items showed low discrimination; of these, eight also were very easy with item p_i-values above 0.80. The test best discriminates at moderate and low ability levels, so making it somewhat more difficult would improve its discrimination among higher ability candidates.

Home Economics Paper 1

Figure 7 below shows the distribution of scores for JC Paper 1 in Home Economics. Table 8 shows a cross-classification of item p_i-values and item r_{it} values for this test.

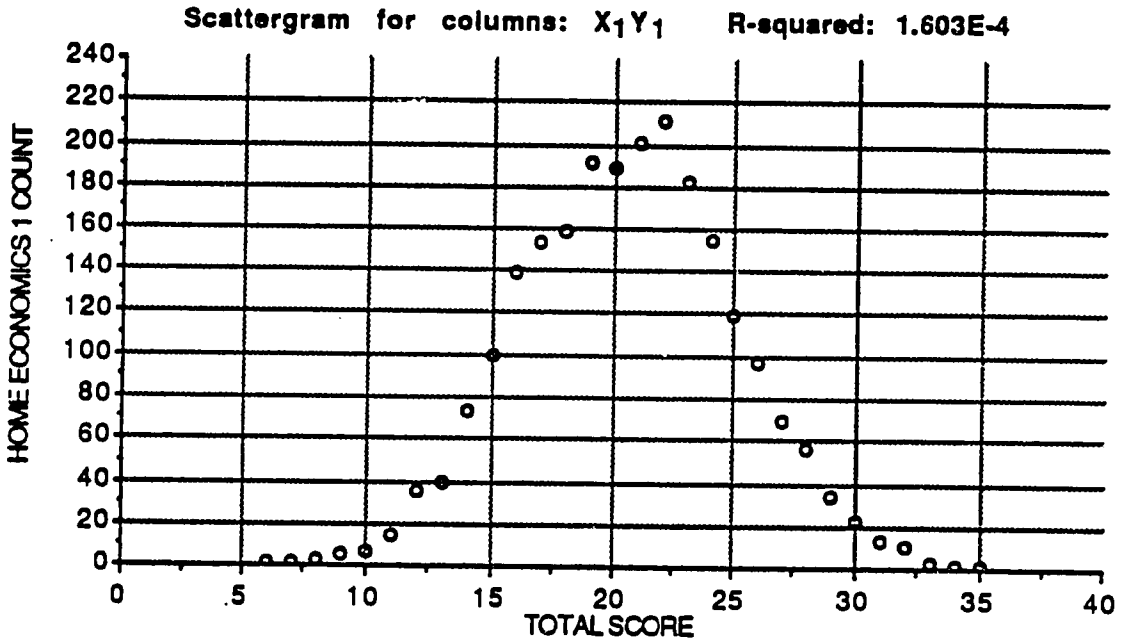

Fig. 7 Distribution of 1987 Two Year JC Paper 1 in Home Economics. (Mean=20.6; S.D.=4.4

Table 8

Cross-Classification of Item p_1 and Item r_{it}
Values for 1987 JC Paper 1 in Home Economics

		Item r_{it} values	
Item p_i	N	N<0.25	N≥0.25
Less than 0.20	2	1	1
0.20-0.39	13	8	5
0.40-0.59	8	2	6
0.60-0.79	12	5	7
0.80 and above	5	4	1
Totals	40	20 50%	20 50%

The Home Economics test has a symmetric distribution with item difficulties spread nicely throughout the range of p_i-values. However, half of its items show low discrimination. This perhaps reflects wider content coverage than that for other subjects. Inspection of these may suggest ways to modify them so that item discrimination and test reliability could be raised.

Religious Education Paper 1

Figure 8 below shows the distribution of scores for JC Paper 1 in Religious Education. Table 9 shows a cross-classification of item p_i-values and item r_{it} values for this test.

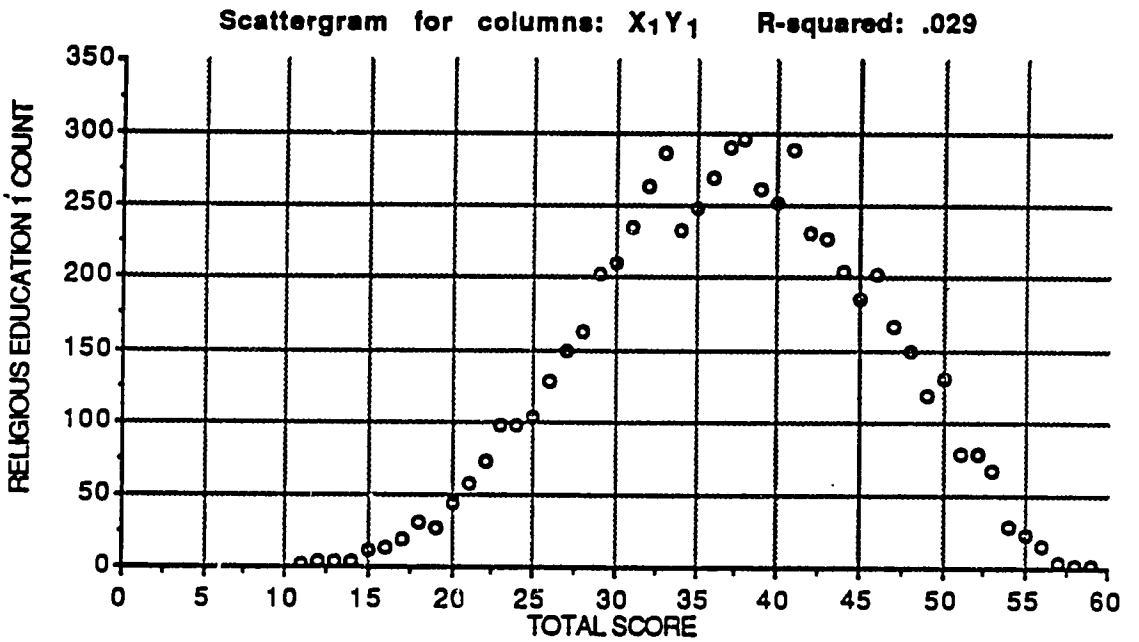

Fig. 8 Distribution of 1987 Two Year JC Paper 1 in Religious Education. (Mean=36.9; S.D.=8.4)

Table 9

Cross-Classification of Item p_1 and Item r_{it} Values for 1987 JC Paper 1 in Religious Education

Item p_i	N	Item r_{it} values N<0.25	N≥0.25
Less than 0.20	2	2	0
0.20-0.39	7	2	5
0.40-0.59	19	4	15
0.60-0.79	21	2	19
0.80 and above	11	5	6
Totals	60	15 25%	45 75%

From a technical point of view, the Religious Education test has excellent features. With 60 items, it is longer than all other tests, so reliability is increased; its spread of item difficulties, aside from too many very easy items, covers a wide ability range; and three-quarters of its items have good discrimination.

Agriculture Paper 1

Figure 9 below shows the distribution of scores for JC Paper 1 in Agriculture. Table 10 shows a cross-classification of item p_i-values and item r_{it} values for this test.

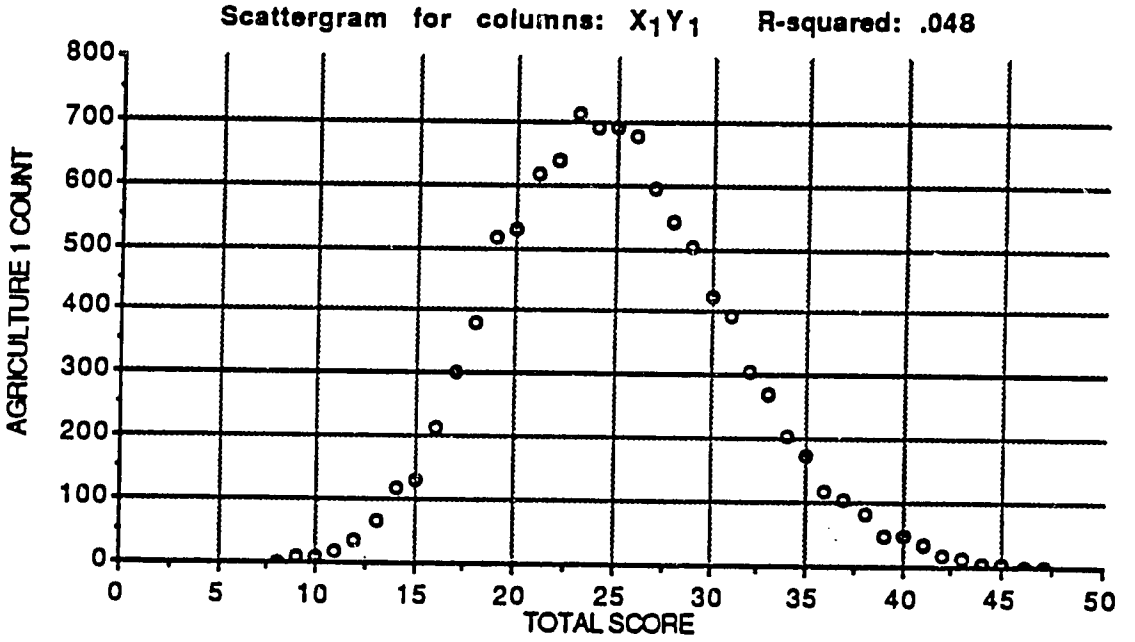

Fig. 9. Distribution of 1987 Two Year JC Paper 1 in Agriculture. (Mean=25.1; S.D.=5.8)

Table 10

Cross-Classification of Item p_1 and Item r_{it} Values for 1987 JC Paper 1 in Agriculture

Item p_i	N	Item r_{it} values N<0.25		N≥0.25	
Less than 0.20	3	3		0	
0.20-0.39	14	6		8	
0.40-0.59	16	3		13	
0.60-0.79	14	3		11	
0.80 and above	3	2		1	
Totals	50	17	34%	33	66%

The Agriculture test has a well-spread distribution of item difficulties. However, the discrimination of difficult items with p_i-values below 0.40 is not entirely satisfactory, with nine of the 17 items showing r_{it} values below 0.25. Also 2 of the 3 very easy items with p_i-values of 0.80 and above show poor discrimination. Revision of items to measure their objectives would lead to better discrimination for both the higher as well as the lower ability candidates.

Summary

Total score and item statistics were reported for nine multiple choice JC Paper 1 tests administered in 1987. Comments were noted for each. The overriding finding is that JC Examinations Sub-Committees in Botswana do an extremely competent job. The concluding comments for this note are:

1. Test construction files or banks containing pretested items that are indexed according to (a) content learning to be tested, and (b) the statistical characteristics of the items, permit test constructors through item selection to pre-specify the means, standard deviations, and reliabilities of JC tests they develop.

2. This enables the development of equivalent tests in a subject from year to year. Equivalent tests in a subject enables changes in candidate achievement to be reliably compared from year to year. With test difficulty now somewhat uncontrolled, "yardsticks" measuring achievement change year to year.

3. **If desired**, statistically equivalent tests across subjects can be developed in the same way. This would enable candidate achievement to be compared from one subject to another. As seen in this report, test distributions now vary from one subject to another. Nevertheless, it is up to each JC Examinations Sub-Committee to decide what test characteristics are intended.

4. Baseline tests can be constructed now at the time when the number of junior secondary schools is rapidly growing. Many students now enter junior secondary school who would otherwise have completed only primary schooling. Baseline tests would enable the Ministry of Education to monitor junior secondary student achievement in future.

What this calls for is not only a regular system of pretesting and item banking, but also a blueprint for test constructors to follow. The blueprint positions instructional objectives to be tested into content categories and levels of learning. It has been used by some JC Examinations Sub-Committees in preparing the multiple choice Paper 1 tests. The blueprint is described more fully in RTC Research Note 88-07.

Combining JC Within-Subject Scores: Nominal and Effective Weighting

Research Note 88-07

July 1988

Research and Testing Centre
Curriculum Development and Evaluation
Ministry of Education
Government of Botswana

Combining JC Within-Subject Scores: Nominal and Effective Weighting

Each JC Examination in a subject consists of a combination of two or three tests (or tests and practicals) that are separately marked. The raw marks are added to a total raw score that is transformed to a standardized score with mean = 50 and standard deviation=15. This is an acceptable procedure, but a problem can arise. We might ask the question: How do the raw marks for the separate tests weight themselves when they are summed to a total raw score? The "nominal" weights are "1" for each raw mark, but what are the "effective" weights? Effective weights, defined here as the relative contribution of each test to the variance of the total raw score distribution, depend upon the distributions and the intercorrelation of the separate marks.

Marks discriminate performance levels among candidates. The variance of a test distribution measures the spread of marks about their average. The larger the spread, the more a test can discriminate performance levels. When marks for two tests are added, the test with the larger variance will affect the variance of the total raw score more than the test with the smaller variance. This note considers the weighting problem when combining only two part scores.

An Example: The 1987 Two-Year JC Art Examination

An example of a JC subject with two parts may illustrate. The 1987 Two-Year JC Examination in Art consists of a multiple choice Paper 1 and a submitted Portfolio. The multiple choice test has a maximum of 40 marks, the Portfolio a maximum of 110 marks. The two marks are added to derive a total raw score.

Statistics for the 1987 Art JC

Table 1 shows the means, standard deviations, and intercorrelations among the 1987 Art JCE Paper 1 marks, the Portfolio marks and the Total Score. A comparison of the standard deviations of Paper 1 (5.6) and the Portfolio (16.9) suggests that the Portfolio outweighs Paper 1 in determining the variance of the Total Score. This is further indicated by comparing the relative correlations of the separate marks with the Total Score -- Paper 1 correlates 0.51 with the Total Score and the Portfolio correlates 0.96. The effective weight for the Portfolio is probably intended to be larger than the effective weight for Paper 1, because a maximum of 40 marks were given Paper 1 and a maximum of 110 marks were given the Portfolio. But are the effective weights close to the intended weights of (40 : 110)?

Table 1

1987 Two-Year Art JCE Statistics

Art Test Score		Mean	S.D.	Intercorrelations (1)	(2)	(3)
(1)	Paper 1	28.1	5.6	1.00	--	--
(2)	Portfolio	64.8	16.9	0.25	1.00	--
(3)	Total	92.9	19.1	0.51	0.96	1.00

Variance = $(S.D.)^2$: Var (1) = 31.36, Var (2) = 285.61.

Effective Weights for Two Raw Part Scores.

Two raw scores (X_1 and X_2) are weighted in the ratio of (W_1: W_2) and then summed to a weighted total score ($W_1X_1+W_2X_2$). VAR (X_1) and VAR (X_2) are the respective score variances, r_{12} is their intercorrelation so that their covariance is $r_{12}\, W_1W_2\, [\mathrm{VAR}\ (X_1)\ \mathrm{VAR}\ (X_2)]^{1/2} = \mathrm{COV}_{12}$.

(1) $[W_1^2\ \mathrm{VAR}\ (X_1) + W_1W_2\ \mathrm{COV}_{12}]$ = the effective weight of W_1X_1.
(2) $[W_2^2\ \mathrm{VAR}\ (X_2) + W_1W_2\ \mathrm{COV}_{12}]$ = the effective weight of W_2X_2.

Conclusion 1. Adding two **weighted raw** scores to form a total weighted raw score gives them nominal weights of $W_1 : W_2$. Their effective weights are in the ratio of $[W_1^2\ \mathrm{VAR}\ (X_1) + W_1W_2\ \mathrm{COV}_{12}] : [W_2^2\ \mathrm{VAR}\ (X_2) + W_1W_2\ \mathrm{COV}_{12}]$.

When two raw part scores are added, the respective weights are (1 : 1). This simplifies Equations (1) and (2) to

(3) $[\mathrm{VAR}\ (X_1) + \mathrm{COV}_{12}]$ = the effective weight of X_1.
(4) $[\mathrm{VAR}\ (X_2) + \mathrm{COV}_{12}]$ = the effective weight of X_2.

Conclusion 2. Adding two **raw** scores to form a total raw score gives them nominal weights in the ratio of (1 : 1). Their effective weights are in the ratio of $[\mathrm{VAR}\ (X_1) + \mathrm{COV}_{12}] : [\mathrm{VAR}\ (X_2) + \mathrm{COV}_{12}]$.

The Art JC. Adding the two part scores of the Art JC gives Paper 1 and the Portfolio nominal weights of (1 : 1). Do these nominal weights yield an effective weighting ratio of (44 : 110) = (1.00 : 2.75)? The effective weights are found from Equations (5) and (6), with figures taken from Table 1. The Portfolio outweighs Paper 1 more than intended, since the effective weighting ratio of (55.02 : 309.27) is approximately (1.0 : 5.5) rather than the intended (1.00 : 2.75).

(5) (31.36) + (23.66) = 55.02 = Paper 1's contribution.
(6) (285.61) + (23.66) = 309.27 = Portfolio's contribution.

Effective Weights For Two Standardized Part Scores.

If we standardize two scores to a common variance -- VAR (X_1) = VAR (X_2) = K, before they are summed to a total score, the effective weights of the standardized scores are, from Equations (3) and (4), shown in Equations (7) and (8).. Both weights are identical, giving an effective weighting ratio of (1 : 1) except in the peculiar case when r_{12} = -1.00.

(7) $K + K\, r_{12}$ = contribution of Paper 1's standardized score.
(8) $K + K\, r_{12}$ = contribution of Portfolio's standardized score.

Conclusion 3. Standardizing two **raw** part scores before they are summed to a total score always gives an effective weighting ratio of (1 : 1) except in the case when r_{12} = -1.00.

The Art JC. The respective nominal weights of (1 : 1) for the two **standardized** scores of the Art JC give an effective weighting ratio of (75 : 75) for the Art JC total score rather than the intended ratio of (40 : 110) = (1.00 : 2.75).

Specifying Effective Weights for Two Raw Part Scores

What should nominal weights for two **raw** scores be to produce an **intended** effective weighting ratio? The answer is found by setting the ratio of Equation (1) : Equation (2) equal to whatever intended effective weighting ratio ($A_1 : A_2$) is desired, setting ($W_1 : W_2$) = ($W'_1 : 1.0$) in both equations, and solving for W'_1 .

(9) $W'^2_1 \, VAR(X_1) + W'_1 \, COV_{12} = A_1$

(10) $VAR(X_2) + W'_1 \, COV_{12} = A_2$

(11) $[A_2][W'^2_1 \, VAR(X_1)] + [A_2][W'_1 \, COV_{12}] = [A_1][VAR(X_2)] + [A_1][W'_1 \, COV_{12}]$,

(12) $[A_2][VAR(X_1)] \, W'^2_1 + [A_2 - A_1][COV_{12}] \, W'_1 - [A_1][VAR(X_2)] = 0$

> **Conclusion 4** If two **raw** scores are to have intended effective weights in the ratio of ($A_1 : A_2$) in a total score, then they should be given respective nominal weights of ($W'_1 : W'_2 = 1.0$). W'_1 is found by solving the quadratic equation: $[A_2][VAR(X_1)] \, W'^2_1 + [A_2 - A_1][COV_{12}] \, W'_1 - [A_1][VAR(X_2)] = 0$.

The Art JC. For the Art JC, intended respective weights = ($A_1 = 1.00 : A_2 = 2.75$). Substituting these and figures from Table 1 in Equation (12) gives Equation (13). Solving for W'_1 gives nominal weights. of ($W'_1 = 1.6$, $W'_1 = 1.0$). The total score which is equal to [(1.6)(Paper 1 raw mark) + (1.0) (Portfollo raw mark)] will be effectively weighted in the intended ratio of (1.00 : 2.75).

(13) $[2.75][31.36] \, W'^2_1 + [2.75-1.00] \, [23.66] \, W'_1 - [1.00] \, [285.61] = 0$

Specifying Effective Weights for Two Standardized Part Scores.

Standardizing to a common variance sets $VAR(X_1) = VAR(X_2) = K$. From Equation (12), we find the root W'_1 in the quadratic equation: $K[A_2] \, W'^2_1 + K[A_2 - A_1] \, r_{12} - K A_1 = 0$, which equals, after eliminating K

(14) $[A_2] \, W'^2_1 + [A_2 - A_1] \, r_{12} - A_1 = 0$.

> **Conclusion 5.** If two **standardized** scores are to have intended effective weights in the ratio of ($A_1 : A_2$) in a total score, then they should be given respective nominal weights ($W'_1 : W'_2 = 1.0$). W'_1 is found by solving for W'_1 in the equation: $[A_2] \, W'^2_1 + [A_2 - A_1] r_{12} \,] \, W'_1 - [A_1] = 0$.

The Art JC. Intended effective weights are ($A_1 = 1.00$: $A_2 = 2.75$). Substituting these and the value $r_{12} = 0.25$ in Equation (14) yields $[(2.75) \, W'^2_1 + (0.4375) \, W'_1 - 1 = 0]$. Solving for W'_1 gives the effective weighting ratio of ($W'_1 = 0.53 : W'_2 = 1.0$). If a total score is equal to [(0.53)(Paper 1 standard score) + (1.00) (Portfolio standard score)], effective weighting will be in the ratio of (1.00 : 2.75) for the Art JC total score.

JCE Overall Merit Ratings Determined from the Average of Seven Subject Standard Scores

Research Note 88-08

July 1988

Research and Testing Centre
Curriculum Development and Evaluation
Ministry of Education
Government of Botswana

JCE Overall Merit Ratings Determined from the Average of Seven Subject Standard Scores

Merit Ratings for Single Subjects. Total raw scores are found for each JC Examination by adding the raw scores of separate tests or practicals. The total raw score is then transformed to a standard score with mean = 50 and standard deviation = 15. Cutoff scores for five merit categories have been set for single subjects. These are listed in Table 1 together with the expected percentage of candidates for each category, when the average standard scores are assumed to be normally distributed with mean = 50 and standard deviation = 15. The expected percentages show that approximately one-fourth of the examinees are expected to achieve either "Good", "Very Good," or "Excellent," about one-half to achieve "Average," and about one-quarter to achieve "Below Average."

Table 1

Expected Percentages for Single JC Subject Merit Ratings Under Current Cutoffs

Cutoff Score	Merit Rating	Expected Percentage
80 and above	Excellent	2.4 %
70-79	Very Good	7.1
60-69	Good	16.9
40-59	Average	49.4
39 and below	Below Average	24.2

Overall Merit Rating as the Average of Seven Subjects. Overall JC merit ratings of candidates are determined from the average of the standard scores earned in seven subjects. Current cutoff levels, expressed as average standard scores for five overall merit categories, are shown in Table 2, together with the percentage of candidates expected in each category under the assumptions that the standard deviation of the average standard score is 15 and that the average standard score has a normal distribution. The expected percentages are similar to those in Table 1. We see that approximately one-fourth of the examinees are expected to achieve either "Merit", "First," or "Second," about one-half to achieve "Third," and about one-quarter to achieve "Unsatisfactory."

Table 2

Expected Percentages for Overall JCE Merit Ratings Under Current Cutoffs

T-Score Cutoff	Merit Rating	Expected Percentage
86 and above	Merit	00.9 %
78-85	First	02.5
60-77	Second	23.1
40-59	Third	49.3
39 and below	Unsatisfactory	24.2

However, this average standardized score is not first restandardized to mean = 50 and standard deviation = 15. Although the mean of the average standard score is 50, its standard deviation is not 15. The standard deviation of the seven-score average is estimated to be approximately 11.3 under the assumption that the mean correlation among the seven scores is 0.50.[1] Under this assumption, it is possible to determine the percentages of candidates expected in the five merit categories when a normal distribution with mean = 50 and standard deviation = 11.3 is assumed. For example, the

cutoff average standard score of 39 for an "Unsatisfactory" merit rating is **not** equivalent to z = [39.5 - 50] ÷ [15] = [-0.70] standard deviations below the mean of 50 but is equivalent to z = [39.5 - 50] ÷ [11.3] = [-0.93] standard deviations below the mean of 50. In this particular example, the expected percentages of candidates with merit ratings of "Unsatisfactory" are 24.2 per cent if the standard deviation is assumed to be 15, and 17.6 per cent if the standard deviation is assumed to be 11.3. For 10,000 candidates, this represents a difference of [2420-1760] = [660] candidates who would be classified as "Third" rather than as "Unsatisfactory." Table 3 below shows a comparison of expected percentages when the standard deviation is taken to be 15 (not correct) and 11.3 (approximately correct).

Table 3

A Comparison of Expected Percentages of Candidates in Five JCE Overall Merit Categories Under Two Assumed Values for the Average Standard Score in Seven Subjects

Specified Cutoff	SD = 15 (Incorrectly assumed) z-score cutoff	Percentage	SD=11.3 (Estimated r=0.50) z-score cutoff	Percentage
86 and above	z= 2.37 and above	00.9	z= 3.13 and above	- - -
78-85	z= 1.83 to 2.37	02.5	z= 2.42 to 3.13	00.8
60-77	z= 0.63 to 1.83	23.1	z= 0.84 to 2.42	19.3
40-59	z= -0.70 to 0.63	49.3	z= -0.93 to 0.84	62.3
39 and below	z= -0.70 and below	24.2	z= -0.93 and below	17.6

Table 4 compares the expected percentages (under both assumed values for the standard deviation of the average standard score) with the actual percentages obtained for the 1987 administration of the 1987 Two-Year JCE. The figures show that the estimated standard deviation = 11.3 leads to expected percentages that are much closer to the obtained percentages for 1987 candidates than does an assumed standard deviation = 15. Cutoff scores set under the assumption that the average standard score's standard deviation = 15 would partly explain the observation that lower proportions than expected of Two-Year JCE candidates earned overall merit ratings of either "Merit, " "First," or "Unsatisfactory."

Table 4

A Comparison of Expected Percentages Shown in Table 3 with Actual Percentage of Candidates in Five JCE Merit Categories

T-Score Cutoff	Merit Rating	Expected if SD=15	Expected if SD=11.3	Actual 1987
86 and above	Merit	00.9 %	--- %	00.1 %
78-85	First	02.5	00.8	00.6
60-77	Second	23.1	19.3	19.2
40-59	Third	49.3	62.3	67.1
39 and below	Unsatisfactory	24.2	17.6	13.0

[1] If the mean and standard deviation for each of K standardized scores are M and SD and the average intercorrelation of the K standardized scores is r, then the mean and standard deviation of the average of the K standardized scores are:

(1) Mean = [K][M]/[K] = [7][50]/[7] = 50.00, when K = 7.

(2) S.D. = [SD][K + K(K-1)(r)]/ [K] = [15][7 + 7(6)(0.50)]/ [7] = 11.3, when K = 7.

A Frame for Writing Examination Items

Research Note 88-09

July 1988

Research and Testing Centre
Curriculum Development and Evaluation
Ministry of Education
Government of Botswana

A Frame for Writing Examination Items

The frame shown in Figure 1 classifies instructional objectives by **content categories** and by **levels of learning.** Content categories are shown horizontally, levels of learning vertically. Instructional objectives are the basis for three educational processes: (1) developing curriculum, (2) constructing examinations, and (3) training teachers. Ideally, all are consistent with the instructional objectives in the frame.

We shall consider four levels of learning. These are arranged in a hierarchy, and each reflects a br capability, described by a "capability verb" indicating what is learned at each level. In general, learning at a lower level is needed for learning at a higher level. We make the distinction between "capability verbs" (verbs that describe unseen learning) and "performance verbs" (verbs that den in a observable and measureable way how a learned capability can be inferred). For example, bein able to demonstrate a rule -- such as finding the sum of two one-digit numbers, is a capability tha can be shown in many ways, but until we specify how this is done in performance terms, we cannot observe the effects of the learned capability.

Levels of Learning	**Subject Content Categories** a	b	c	d	e	**Capability Verbs**
Higher Abilities:			O_1			Analyse, Originate
Rules: Applications		O_2				Solve, Demonstrate
Concepts: Comprehension				O_3		Identify, Classify
Facts: Basic Knowledge					O_4	State, Name

Instructional Objectives in the Frame

Four instructional objectives are symbolically shown in the frame as O_1, O_2, O_3, and O_4. An instructional objective may be expressed in the following way:

> Given (**specified conditions**), the student will (**show a learned capability**) by (**performing a specified action**), resulting in (**a specified product**).

This statement of an instructional objective (1) distinguishes between learning and performance, (2) does not specify a standard of performance (or criterion) that enables one to evaluate and to ir the extent to which a capability has been learned (we reserve statements of criteria for item writin specifications), and (3) indicates what examinees must do to enable us to infer whether the learnin specified by the instructional objective has been acquired.

Item Writing Specifications

Test items are based on instructional objectives. They are indicators that enable us to determine whether intended instructional objectives have been achieved. Any single instructional objective may form a basis for many tests items or for different types of test items. For example, we could ask an examinee to show whether he or she has learned the rule for adding two two-digit numbers by presenting: "71 + 32 = ?", by asking "The sum of 95 and 14 is ?", or by a word problem. The item writing specification tells item writers what behaviours a test item measures, and details the behavioural domain so that they can write functionally similar items.

Test items may be broadly classified either as those that require the examinee to **select** a response or those that require the examinee to **construct** a response. A multiple-choice item or a matching item are examples of selection items, a short-answer question or a question asking for an essay are examples of **construction** items. Whatever the case, the item writing specification includes:

(1) The **instructional objective** on which it is based.

(2) The **stimulus elements** of the item; these indicate the content covered by the item, and the conditions and constraints that are to be considered by the item writer.

(3a) The **response standards** for selection items; these are rules given items writers for preparing selection responses that specify a correct answers and right-wrong options.

(3b) The **response standards** for construction items; these are rules for constructed responses that must clarify in detail what constitutes an appropriate answer.

(4) The **criterion** or rule for marking correct and incorrect answers.

Sample Items. A sample item is written that includes the **instructional objective**, the **stimulus elements**, the **response standards**, and the **criterion.** An example follows.

Instructional Objective. Given a set of six two-digit numbers, the student will demonstrate that the concept "odd number" has been learned by identifying the odd numbers in the set and drawing a circle around them.

Stimulus Elements. Present a set of six two-digit numbers. Two are odd and four are even. The odd numbers end with 1, 3, 5, 7, or 9. The even numbers end with 2, 4, 6, or 8. One of the odd numbers starts with an even number, the other starts with an odd number. Two of the even numbers start with an even number, two with an odd number. The six numbers are presented as a row of six numbers in any order.

Response Standards and Criterion. Students will circle numbers they choose as odd with a pencil. Responses that are changed must be cleanly erased. The student receives one mark for each odd number that is properly circled. The student is given a penalty of one mark for each even number that is circled. Maximum mark is 2; minimum mark is -4.

Item A. Six numbers are shown below. **Circle the odd numbers** with your pencil. Carefully erase any mistakes. You receive one mark for each odd number that you circle, but you will lose one mark for each even number that you circle.

43 54 28 82 96 37