PN-ABU-940

REPORT ON THE TRAINING WORKSHOPS ON TEST ANALYSIS AND INTERPRETATION AT LAHORE, HYDERABAD, AND QUETTA

PRESENTERS:

DR. SAR BILAND KHAN

DR. KRISTIAN JOHN KIRKWOOD

AED CONSULTANTS

WASHINGTON, D.C.

APRIL 1994

EXECUTIVE SUMMARY

The purpose of each workshop was to 1) to review the basic principles of measurement related to test analysis; 2) to demonstrate the utility of item analysis programs; 3) to introduce the participants to computer programming for item analysis and secondary analyses1.

A total of 45 individuals participated in the three workshops (Lahore 21, Hyderabad 11, and Quetta, 13).

Even though the same material was covered in all three of the workshops, the topics were not all covered in the same sequence because the needs in each of the provinces were unique. Correspondingly, a separate summary is presented for each of the three workshops.

The workshops included the following topics: the decision-making process, types of decisions made in education, the current testing movements in Pakistan, obtaining accurate ad relevant information, assessing academic achievement, basic statistics, using the item analysis program. interpretation of item and test results, and finally using test results for diagnostic

purposes.

Overall, according to the workshop participants, the workshops were worthwhile and provided them with the essential skills to interpret test results.

TABLE OF CONTENTS

Executive Summary ii

Lahore: March, 19-22, 1994 1

Hyderabad March 26-29, 1994 8

Quetta April 5-7,1994 14

Summary 19

Recommendations 21

Appendix A: Workshop Outline 24

Appendix B: Name of Participants 26

Appendix C: Issues in Testing 29

Appendix D: Glossary of Terms 51

Appendix E: Item Analysis Printout 56

Appendix F: Steps in Field Testing 65

Appendix G: Levels of Measurement 72

Appendix H: The Need for Standardization of Raw Scores . . 73

Appendix I: Percent Success-Rate on Math Items for NWFP 77

Appendix J: Percent Success-Rate on Math Item for Balochistan 111

Apppendix K: Summary of Meeting in Islamabad with Mr. Abbassi Co-Joint, Secretary, Primary Education Directorate, March xx, 1994 132

Appendix L: Minutes of the Meeting with IBA, University of Karachi, March 31, 1994 133

LIST OF TABLES

Table 1: Summary of Participants by Gender and City 23

Table 2: Computer Resources Available by City 24

LAHORE

DAY 1: MARCH 19, 1994first day of the workshop began with a brief introduction by Ms. Yasimin Gul Qamar, Additional Director, Curriculum Research and Development (CRDC) followed by recitation from the Holy Quran. On the first day, the workshop was attended by 20 participants. (See Appendix B for the names and positions). Dr. Sar B. Khan initiated the presentation by outlining the objectives of the workshop. These included:

1) to review the basic principles of measurement related to test analysis;

2) to introduce the participants to computer programming for item and test analyses as well as secondary analyses;

3) to demonstrate the utility of item analysis and test analysis in instructional diagnosis.

In addition to listing the objectives, Dr Khan emphasized that this workshop was not on test development *per se* but mainly on how to analyze and interpret test responses and how

to use information from tests in educational decision-making, i.e., improving the quality of instruction.

Dr. Khan briefly described the decision-making process and engaged the participants in listing situations in which each of them was required to make a decision that morning. The idea of the existence of alternatives/choices was extracted from the discussion. The presentation moved to the need for information (accurate and relevant) in order to be able to make probably correct decisions. The presentation was linked to decision-making in education such as instructional, promotion/certification, selection/admission, administrative/managerial, etc. supported by examples for each of the above situations. Various information-gathering procedures ranging from asking a simple question to the administration of tests and examinations in a formal setting were discussed. During this discussion, the need for adequate number of relevant observations was emphasized in order to be able to draw reliable and valid inferences from the information collected.

Dr. Kris Kirkwood introduced the item analysis program for scoring and analysis and outlined the salient features of the

program. It was noted that the item analysis program could score 30 000 examinees in one file and could analyze up to ten subtests. He showed the participants how write the format statements and the type of statistics provided by the item analysis program. A sample handout of a test analysis using the item analysis program was given to the candidates and the contents were reviewed.

In the next session, participants were introduced to the computer. It was interesting to note that not a single participant had any experience with computers. The participants were given the rudiments of how to turn the computer on and how to activate the item analysis program.

At the end of the session, the participants were given reading materials in the form of handouts and were requested them to read the materials before the second day of the workshop.

DAY TWO: MARCH 20, 1994

The morning started with a recitation from the Holy Quran. This was followed by a review of the salient concepts which were introduced on the first day. Students were given the opportunity to ask questions for clarification or for additional information. The purpose of this exercise was to ensure that all participants understood the basic statistical and measurement concepts.

Next, the participants were instructed on the how to run the item analysis program. Each of the three groups (mathematics, science, and Urdu) were asked to run the item analysis for their respective subject area. Each group wrote their format statements for the item analysis program. Then each student was given the opportunity to actually run the program and see the type of output that the program generated. The data utilized for this exercise were provided by the NWFP from their recent testing program (NEAP).

The students then conducted an item by item analysis for each test. The first part of the analysis consisted of identifying the key concepts that were purported to be measured on each

test. This was followed by the frequency tables including p-values and biserials. In addition, the participants constructed a table based on the scale statistics - mean, median, standard deviation, skewness, range, mean biserial, and Cronbach's alpha.

The highlight of the day from the participants perspective was the time spent on the computers. The participants really appeared to enjoy the computer assignment since it was the first time that they actually operated a computer in a meaningful way. A number of participant revealed that "this was a truly great day" because of the initiation to computing.

At the end of the session, additional handouts were given to the participants and they were asked to read them in preparation for the next session.

DAY THREE: MARCH 21, 1994

The morning started with a recitation from the Holy Quran. Dr. Khan started the morning by stating that the three groups should work in small groups and complete the item analysis. Model tables were given to each group and they were requested

to make similar tables based on their data.

The groups were re-assembled and a discussion of their respective findings ensued. The groups were asked to make individual reports as to their opinions on the concepts which were the easiest and the most difficult for the examinees. Based on the participants' remarks, a number of concepts related to the item analysis were re-visited in order to ensure that all of the participants truly understood them.

For the final portion of the day, the students returned to the computer for additional practice.

DAY FOUR: MARCH 22,1994

The morning started with a recitation from the Holy Quran. Once again the key concepts related to item analysis were reviewed and the students were asked if there were any additional points which needed clarification. Then Dr. Khan discussed in detail the interpretation of each of the statistics contained in the scale statistics generated by the item analysis program.

Dr. Kirkwood administered a test based on 10 items that

dealt with measurement concepts. The purpose of this exercise was to allow the participants an opportunity to generate their own data and then enter it on to the computer after they had written the format statements. The participants entered their own data and also entered the format statements. The program was run and the output was run off in multiple copies. The students then analyzed the test based on their acquired knowledge of item analysis.

The workshop ended with Ms. Yasimin Gul Qamar thanking the presenters and indicating that the workshop was long overdue given the current emphasis on testing in Pakistan. She stated that the participants revealed that the felt that the workshop was informative and worthwhile. They were looking forward to future workshops in item and test analysis to expand and re-enforce what they have learnt in this workshop.

HYDERABAD

DAY ONE: MARCH 26, 1994

The workshop started with a recitation from the Holy Quran. Mr. Ranjho Khan Zardari then introduced the workshop presenters and Dr. Sar Biland Khan outlined the objectives of the workshop. These included:

> 1) to review the basic principles of measurement related to test analysis;
>
> 2) to introduce the participants to computer programming for item and test analyses as well as secondary analyses;
>
> 3) to demonstrate the utility of item analysis and test analysis in instructional diagnosis.

Each participant was asked to introduce themselves indicating their respective position and the organization for whom they worked. There were a total of 11 participants - ten males and one female. (See Appendix B for the names and positions).

Dr. Khan then began with a systematic review of the decision-making process and emphasized the importance of

having accurate and relevant information so that the best possible alternative could be selected. Decision-making in the educational context was reviewed next. Some of the types of decisions that educators make include instructional, promotion, certification, admission, and administrative. It was concluded after this discussion that in order to obtain valid and reliable information, there was a need for an adequate number of accurate and relevant observations.

DAY TWO: MARCH 27, 1994

The morning started with a recitation from the Holy Quran. Then Dr. Khan reviewed the salient points that were discussed on day one. This was followed by an indepth discussion of some the issues and concern associated with testing in Pakistan. These included the lack of standardization of curriculum, teacher training problems, frequent incidents or irregularities and malpractice and why information from the current examination is least useful in educational decision-making.

The workshop then shifted from the discussion to the

practical work of learning about the computers and the item analysis program. Dr. Kirkwood introduced the basic elements of computing and then instructed the participants on how to prepare the format statement for the item analysis program. Because the Bureau of Curriculum computer department had SPSS version PC, the participants were also instructed in the use of SPSSPC. This included the use of the SPSS text editor "REVIEW" and how to make and edit data files. In addition, the participants were shown how to write program statement for secondary analysis of the data emanating from the item analysis. The participants were very interested in this component and they supplied some the test results. These results originated from one teacher training college in Sindh. The test item responses were entered on to the computer for item analysis and subsequent analysis utilizing SPSSPC. It was evident that having **real** data heightened the interest level of the participants.

DAY THREE: MARCH 28,1994

Day Three started with a recitation from the Holy Quran.

This was followed by a brief discussion where the class elected to spend more time working on the computers. All participants were given the opportunity to run the item analysis program and obtain a printout of the results.

After each of the participants received their printouts, then Dr. Khan went over in detail what each of the different components of the printout mean. Simultaneously, he referred to each of the different handouts which added additional information to the analysis. The day concluded with a lengthy discussion of what the actual statistics actually mean. It was decided that the last day would start with a comprehensive review of the test analysis.

DAY FOUR: MARCH 29, 1994

The session started with a recitation from the Holy Quran. Then Dr. Kirkwood asked the participants to retrieve three handouts from their files for the review of the test analysis. These included: **Issues in Testing, Glossary of Terms**, and the **printout** from the Sindh mathematics test from the one teacher

training college. Each column of the printout was reviewed and the corresponding sections of the **Issues in Testing** were identified so the participants could use it for future reference. This process was done in a step by step method until all participants were able to understand the concept.

Then, the participants were asked to review the concepts that the test purported to measure. This was done for each item. The item difficulty level was also examined and participants were asked to identify items that most students did not successfully complete and also to identify items that were considered relatively easy by the students. The cut off for difficulty was 30% or less and for easy items it was 80% and above. Based on this analysis, the participants were requested to suggest some instructional changes for the college program for each of the subject areas - Urdu, science and mathematics.

The workshop ended with Mr. Zardari thanking the presenters. He noted in his concluding remarks that he felt everyone learned a great deal and he was sure that they have reached a new plateau in test analysis. The students individually

thanked us and stated that this was, indeed, a very useful and informative experience. They indicated that they would like additional workshops in the immediate future.

QUETTA

DAY ONE: APRIL 5:1994

The session opened with a recitation from the Holy Quran. This was followed by an introduction of the presenters by Miss Sarwat Ismail. There were a total of 13 participants - eight females and five males. (See Appendix B for names and positions). Dr. Khan outlined the purpose of the workshop and presented the objectives of the workshop. These included:

1) to review the basic principles of measurement related to test analysis;

2) to introduce the participants to computer programming for item and test analyses as well as secondary analyses

3) to demonstrate the utility of item analysis and test analysis in instructional diagnosis.

The workshop focussed on the basic principles associated with the decision-making process. The importance of possessing data which is both accurate and relevant was highlighted. Participants were invited to discuss their views on the decision-

making process e.g., how they perceived decision-making within the context of their own jobs. Next, the types of decisions that educators make on a daily basis were considered. Decisions about promotion, certification, admission, and administration /management were discussed. In summary, the participants agreed that there was a need to obtain information that was accurate and relevant so that the **good** decisions can possibly be made.

Next, the students were introduced to the item analysis program. The basic principles of how to invoke both the item analysis program and the text editor for SPSSPC. Students then completed an assignment where they assembled the four format statements which are required for the item analysis program. Each participant had the opportunity to work independently on one of the four computers.

DAY TWO: APRIL 6, 1994

After the recitation from the Holy Quran, the participants were divided into four groups (one group per computer and they

spent the entire session developing programs to analyze the field test data from six Baluchistan colleges in Urdu, science and mathematics. Before embarking on this task, the key statistical concepts associated with item and test analyses were reviewed. (See Appendix C). The groups worked on writing the format statement for the item analysis program and then they obtained their outputs for their respective subject areas - Urdu, mathematics, and science. An example of the item analysis output is given in Appendix E. Each participant was responsible for obtaining a printout for his/her personal files. When each participant had completed this task, the groups were asked to identify the major concepts that each item purported to measure. Then, they were asked to identify items that most students completed successfully (easy items) and items that the majority of student did not successfully complete (difficult items). (See Appendix H for an example of the results of this exercise). Finally, the participants were requested to make diagnostic and instructional recommendations based on the results of the three tests.

DAY THREE: APRIL 7, 1994

The morning started with a recitation from the Holy Quran. Because of the facilities provide (four computers), we were able to cover more material in a shorter period of time. Consequently we were able to instruct the participants in how to perform secondary analyses on the data generated by the item analysis program. The participants were shown how to write SPSSPC programs including frequencies, t-tests, correlations, and analysis of variance. Each group performed a t-test to determine if there were any gender differences between the scores on the various tests. Similarly, programs were written to examine differences in scores by academic qualifications and college.

For the final session the results of the secondary analysis were discussed and the participants gained some insight in the meaning of statistical significance.

Miss Sarwat Ismail, organizer of the training workshop, thanked the presenters stating that the duration was too short and she would have liked us to stay even longer. She noted that the enthusiasm generated by the participants was a manifestation

not only of their needs in this area but also of the positive manner in which the workshop was presented.

SUMMARY

Based on our personal experiences and the comments from the organizers of the three workshops, it is concluded that the workshops were successful. The participants both individually and collectively stated that they learned a great deal in a very short period of time. They lamented, however, that the time was too short and they would like similar workshops in the immediate future that re-enforce their newly acquired skills in the area of item and test analyses, and to further develop skills in how to perform secondary analysis and how to interpret the results of such analysis.

Overall, there were very few participants who had ever worked on a personal computer. Besides learning about item and test analysis, they truly appreciated the opportunity to actually work on a computer. The workshops provided this opportunity. It is of interest to note that when the daily sessions were completed many participants elected to stay and spend more time

on the computers - a testament to their enthusiasm and interest. There is no doubt that more resources in terms of computers would have facilitated the workshops, however, the participants were very magnanimous in sharing their time on the computers.

RECOMMENDATIONS

Given the meteoric rise of the status of testing in Pakistan in recent months, it is important to inform officials at all levels of the educational hierarchy about the need and utility of having psychometrically sound testing programs both at the provincial and national levels. In addition, it is critical that individuals be taught how to construct and develop valid and reliable tests if the testing program are to be credible and transparent.

Based on the experiences in the three workshops, we offer the following recommendations:

> *1) that a second advanced series of workshops be conducted in order to enhance the skills acquired as a result of this workshop. Unless these types of technical skills are periodically re-enforced the individuals will lose what information and skills they acquired in these workshops.*
>
> *2) that future workshops of this nature have access to a least one computer for every two or three participants.*
>
> *3) essential software for the test analysis and statistical analysis should be available*

on all computers.

4) supporting materials such as computer paper, diskettes, ribbons, etc. should be readily available to the users.

5) computers should be accessible for use to every staff member of the organization in which they are housed.

6) adequate facilities such as air conditioning, ventilation, electrical outlets, stabilizers, etc., should be provided in the areas in which the computers are housed.

It is axiomatic that no one workshop can offer a panacea to the myriad of problems associated with testing in Pakistan. Nevertheless, it is only through systematic efforts that these problems can be eradicated over time. These types of workshops certainly make an impact on two levels - awareness of the problems and the development of technical skills. In order to ensure the change evolves in the appropriate direction a sustained effort must be made.

Table 1: Summary of Participants by Gender and City

Gender	Lahore	Hyderabad	Quetta
Female	7	1	8
Male	14	10	5
Total	21	11	13

Table 2:Computers Resources Available by City[1]

City	Computers	Printers	Wordprocessing
Lahore	3	2	Mircosoft Word 5
Hyderabad	2	2	Microsoft Word 5
Quetta	4	2	Wordperfect 5.1

[1]This does not include the laptop that we carried with us to each site.

APPENDIX A:

WORKSHOP OUTLINE

1. Recitation from the Holy Quran and Introductions
2. Objectives of the Workshop
3. The Decision-making Process
4. Types of Decisions Made in Education
5. Background to the Current Testing Movement in Pakistan
6. How We Obtain Information
7. Assessment of Academic Achievement Learning
8. Criterion-referenced Measurement
9. Levels of Measurement
10. What Do We Do with the Information/Data
11. Review of Basic Statistics
12. Item and Test Analyses
13. Use of Computers for Item and Test Analyses
14. Interpreting Results from Item and Test Analyses
15. Use if Item and Test Analyses for Diagnostic Purposes

APPENDIX B:LIST OF PARTICIPANTS

WORKSHOP PARTICIPANTS
LAHORE, PUNJAB
MARCH 18-23,1994

	POSITION	INSTITUTION
RASHIDA TALMIZ	RESEARCH ASSOCIATE	CRDC
NASIM SABIRA	RESEARCH ASSOCIATE	CRDC
ATIQA AMJAD	RESEARCH ASSOCIATE	CRDC
MRS. SAIYED	RESEARCH ASSOCIATE	CRDC
FARHAT SALEEM	RESEARCH ASSOCIATE	CRDC
MOHAMMAD ASHRAF NAZ	DEPUTY DIRECTOR	CRDC
BASHIR AHMED ARSHI	DEPUTY DIRECTOR	CRDC
IFFAT RIZIR	RESEARCH ASSOCIATE	CRDC
MUHAMMAD ASHRAF	DEPUTY DIRECTOR	CRDC
CHIRAGH DIN ARIF	RESEARCH ASSOCIATE	CRDC
MOHAMMED NAZIN	RESEARCH ASSOCIATE	CRDC
MOHAMMED AFZAL	RESEARCH ASSOCIATE	CRDC
GOLULAN ABBAS SARGANA	DEPUTY DIRECTOR	CRDC
ZAHOOR ANWAR	DEPUTY DIRECTOR	CRDC
GUL YASMIN OAMAR	ADDITIONAL DIRECTOR	CRDC
MEHTAB ALI KHAN	RESEARCH ASSOCIATE	CRDC
JAVAID ALI CHANDRY	DEPUTY DIRECTOR	CRDC
SOFIA SULTANIA	SUBJECT SPECIALIST	DSD
SHAMIM RIZIR	RESEARCH OFFICER	DSD
MOHAMMED MUNIR	SUBJECT SPECIALIST	DSD
MOHAMMED JAMIL CHANDRY	SUBJECT SPECIALIST	DSD

WORKSHOP PARTICIPANTS
HYDERABAD, SINDH
MARCH 26-29, 1994

NAME	POSITION	INSTITUTION
RANJHO KHAN ZARDARI	SUBJECT SPECIALIST	BUREAU OF CURRICULUM
MUHAMMED SHARIF KALHORO	SUBJECT SPECIALIST	BUREAU OF CURRICULUM
GHULAM AKBAR KAZI	RESEARCH OFFICER	TEACHER TRAINING PROJECT - ADB
JANVARIUS PARVAIZ	PRINCIPAL	GITTC
ABDUL RASHEED ARAIN	STATISTICAL OFFICER	BUREAU OF CURRICULUM
GHULAM ASGHAR MEMON	SUBJECT SPECIALIST	BUREAU OF CURRICULUM
ISHRAQ UDDIN AHMED	INSTRUCTOR	GECE(M)
ABDUL MAJEED BHURT	SUBJECT SPECIALIST	BUREAU OF CURRICULUM
RASOOL BISX SAMOO	REGISTRAR	BUREAU OF CURRICULUM JAMSBORO
ABDUL HAFEEZ MEMON	AUDIO VISUAL AID OFFICER	BUREAU OF CURRICULUM
ZEENAT ALLANA	LECTURER	ELEMENTARY COLLEGE OF EDUCATION

WORKSHOP PARTICIPANTS
QUETTA, BALOCHISTAN
APRIL 5-7, 1994

NAME	DESIGNATION	INSTITUTION
NUZHAT GULZAR	SUBJECT SPECIALIST	BIMDTC
SARWAT ISMAIL	TESTING CONSULTANT	BUREAU OF CURRICULUM
NOSHABA MEHMOOD	SUBJECT SPECIALIST	BUREAU OF CURRICULUM
TAHIRA QAZALBASH	ASSISTANT DIRECTOR	BIMDTC
MUMTAZ ANWAR	SUBJECT SPECIALIST	BUREAU OF CURRICULUM
NAJMA WAHID	SUBJECT SPECIALIST	BUREAU OF CURRICULUM
FAKHRA RAMZAN	PROGRAM ASSOCIATE	BIMDTC
M. MAHMOOD ATIQUE	SUBJECT SPECIALIST	BUREAU OF CURRICULUM
MOHAMMAD ASGHAR	SUBJECT SPECIALIST	BUREAU OF CURRICULUM
MAHMOOD ATIQUE KHAN	DEPUTY DIRECTOR	BIMDTC
ABDUS SATTAR MUJAHID	LECTURER	UNIVERSITY OF BALUCHISTAN
ABDUS SATTAR	SUBJECT SPECIALIST	BIMDTC
ERUAM FAROOQI	PROGRAM ASSOCIATE	BIMDTC

APPENDIX C:

TECHNICAL ISSUES

BASIC STATISTICS
ITEM/TEST ANALYSIS
TEST RELIABILITY
TEST VALIDITY
INTERPRETING TEST RESULTS

Sar B.Khan
AED Consultant
Washington, D.C.

February, 1994

I. BASIC STATISTICS[1]

In order to understand the material in the following section, knowledge of some basic statistics is essential. Statistics deals with information given in the form of numbers. As test scores are numbers, we can use various statistical techniques to analyze and summarize test-score data. The statistical concepts we will be dealing with for the purpose of understanding the material in the following section are: mean, variance, and correlation. [[1] Glossary of essential terms is given in Appendix D]

(i) Mean (Average)

Let us say that you have administered a ten-item multiple-choice test to 10 individuals and the following scores are obtained.

Person Code	A	B	C	D	E	F	G	H	I	J
Test Score	8	8	4	4	5	6	7	3	3	2

In actual practice, we will have a large number of scores obtained from a test containing a large number of items. The question is how could we summarize the test-score data to make meaningful interpretations about the performance of the students. Statistical index called mean or the average provides information on how did the students do on the test as a group. For obtaining, the mean or average of the score, all scores are added and then the sum is divided by the number of individuals in the group. In our example, the sum of the scores is 50. The number of individuals in the groups is 10. Therefore, the mean score of the group is 5. Specifically, the formula for obtaining the mean is:

$$\overline{X} = \frac{\Sigma X}{n}$$

where, $\overline{X}$ = mean;

X = test score of ar individual;

Σ = (sigma); summa ion sign meaning that all scores (X's) are added;

n = number of indiv duals in the group.

(ii) Variance

Another statistical index called variance provides information on the spread in the scores. If scores of 100 students on a 50-item range from 8 to 48 compared to another group of 100 students whose scores range from 31 to

45, then there is more variability in performance within the first group in comparison to the second group. Although the above method (computing range) tells us something about the spread of the test scores, it does not utilize the information based upon all test scores. The index which is based upon all of the test scores is called variance.

Specifically, variance is given by

$$\frac{\Sigma (X-\bar{X})^2}{n}$$

The following steps are involved in obtaining variance.

1. Subtract the mean ($\bar{X}$) from each score (X).
2. Square all the differences. There will be as many differences as the number of students in the group. Some differences will be negative (the score is less the mean), some will be positive (the score is above the mean), and some differences will be zero (if the scores are exactly equal to the mean).
3. Sum all the squared differences and divided by n to obtain variance. We should be using (n-1) rather than n in the denominator, however, for all practical purposes it will not make much difference in the value especially, if the n is quite large.

Let us obtain the variance in our example. If we subtract the mean from each score, square the differences, then the sum of the squared difference will be 42. Dividing 42 by 10, we get 4.2 as the variance. In other words,

$$\Sigma (X - \bar{X})^2 = 42$$

$$n = 10$$

$$\therefore \text{ variance (V)} = \frac{42}{10} = 4.2$$

As you know, we had squared the differences in order to obtain variance. In order to bring the index of variability to the same scale as the test scores, we will have to take the positive square root of the variance. This index is called standard deviation. In our example,

$$\text{Standard Deviation (S)} = \sqrt{4.}$$

Both mean and variance are required to fully describe the distribution of test scores. While the concept of average is not very difficult to understand, the concept of variance may be somewhat difficult to understand. Think of variance as an index of homogeneity or similarity of test performance. The larger this value, the less homogeneous or more dissimilar the performance of the individuals. The smaller this value, the more homogeneous or less dissimilar the performance of the individuals. For instance, if we obtain variance of test scores (test containing 50 items) for 100 boys to be 121 while the variance for 100 girls (the same test) is 64, then boys as a group are more variable (dissimilar in performance) than the girls.

(iii) Correlation

There are times when we have more than one test score for each student in the group. We may like to see if there is any relationship between the two sets of scores. The statistical index to use in this case is called a correlation coefficient. Correlational analysis is very useful in understanding association between two variables and predicting performance on one variable from the knowledge of performance on the other variable.

A correlation coefficient indicates the degree of association or relationship between measures on two variables. It ranges in value from -1.0 (perfect negative correlation) to +1.0 (perfect positive correlation). Graphic representation of different magnitudes of correlation is given in Appendix D. A correlation of zero indicates complete absence of relationship between two variables. For instance, one would expect a positive correlation between physics test scores and mathematics test scores. On the other hand, one would expect a negative correlation between supply of a commodity and its price when the demand is kept fixed. In other words, as the supply increases, price will decrease and as the supply decreases, price will increase. One will not expect any correlation (a correlation of around zero) between length of the right arm and performance on a mathematics achievernent test. The reason is that there is nothing in common between how long or short one's arm is and how well one performs on a mathematics test.

Suppose you have accumulated scores on a spelling test (X) and a verbal comprehension test (Y) for 20 students. Now you have 20 paired scores (Hameeda's score on the spelling test and her score on the verbal comprehension test is one pair) which you can manipulate statistically and

compute a correlation coefficient.

In order to compute a correlation coefficient between a set of paired data, we require information on the mean and variance of each test. Let us examine the example given in Table 1.

Table 1
Example of Calculating a Correlation Coefficient

	X	Y	X - $\overline{X}$ (x)	(y)	Y - $\overline{Y}$	xy	x^2	y^2
	5	6	0	+1	0	0	1	
	4	3	-1	-2	2	1	4	
	2	2	-3	-3	9	9	9	
	1	1	-4	-4	16	16	16	
	6	7	+1	+2	2	1	4	
	7	7	+2	+2	4	4	4	
	9	7	+4	+2	8	16	4	
	6	5	+1	0	0	1	0	
	8	6	+3	+1	3	9	1	
	4	3	-1	-2	2	1	4	
	3	4	-2	-1	2	4	1	
	6	5	+1	0	0	1	0	
	2	4	-3	-1	3	9	1	
	6	6	+1	+1	1	1	1	
	6	7	+1	+2	2	1	4	
	5	6	0	+1	0	0	1	
	4	3	-1	-2	2	1	4	
	6	5	+1	0	0	1	0	
	5	6	0	+1	0	0	1	
	5	7	0	+2	0	0	4	
Totals	100	100	0	0	56	76	64	

$\Sigma X = 100$ $\Sigma Y = 100$

$n = 20$ $n = 20$

$$\overline{X} = \frac{\Sigma X}{n} \qquad \overline{Y} = \frac{\Sigma Y}{n}$$

$$= \frac{100}{20} \qquad = \frac{100}{20}$$

$= \quad 5.00 \qquad\qquad = \quad 5.00$

$$\text{Variance} = \frac{\Sigma x^2}{n} \qquad \text{Variance} = \frac{\Sigma y^2}{n}$$

$$= \frac{64}{20} \qquad = \frac{100}{20}$$

$$= 3.8 \qquad = 3.2$$

$$S_x = \sqrt{3.8} = 1.9 \qquad S_y = \sqrt{3.2} = 1.8$$

Let r denote correlation. The formula for r is

$$\frac{\Sigma\, x\, y}{n\, S_x\, S_y}$$

Now $\Sigma xy = 56$, $S_x = 1.95$, $S_y = 1.80$, $n = 20$

$$\therefore r = \frac{56}{20 \times 1.95 \times 1.80} = \frac{56}{70.2} = .80$$

We should have (n - 1) rather than n in the denominator. However, for convenience, we have used n in the denominator.

Remember that correlation can only be obtained on several paired measurements, that is, measurements on two variables (or the same variable at two points in time) on the same group. There has to be an adequate number of measurements for the correlation coefficient to be meaningful. Generally speaking, the larger the number of measurements (individuals in a group on whom measurements are taken), the more stable or accurate the correlation.

Correlation between the two measurements on the same individuals can exist only if there is variation on both measurements. The correlation will generally be small if there is less variation on one or both measurements. Let us consider an example. You have given a 20-item test to 100 students covering multiplication to 10 (that is from 2 to 10; the last being 10 x 10), and you obtained scores ranging from 0 to 20. Then you gave them sufficient drill on the multiplication tables during a two-week period. You administered the same test again and obtained scores ranging from 16 to

20. The correlation between the two sets of scores will be quite small (almost non-existent) because you have reduced variation in the performance of students as indicated by the posttest scores.

II. ITEM ANALYSIS

After a test is scored, it is important to examine each item in terms of some relevant criteria. This is very important if the items are to be used again or they are to become part of an item bank. One can evaluate an item in terms of the information it provides on the status of performance of the individuals. In other words, if an item contributes to our knowledge of individual differences (when one of the purposes of the test is to assess individual differences rather than mastery in a certain domain of content) in a meaningful way, then it is worthwhile to keep that item for future use. Item analysis is defined as the process of examining responses to each item in terms of its **difficulty level** and **discriminating power** as well as the effectiveness of each of the incorrect options (distractors).

As indicated earlier, item analysis can be carried out by the computer as part of scoring and analyzing test responses (see an example of an item analysis by computer of a 30-item test given to 400 students in Appendix E).

(i) Difficulty Level

Difficulty of an item is defined as the percentage of examinees passing (being successful) on an item. The higher the difficulty index in terms of magnitude, the easier the item; the lower the difficulty index, the more difficult the item. Very easy items and very difficult items do not tell us much about individual differences. It is suggested that the majority of items should be in the average difficulty range. It is a good practice, however, to have a few easier items at the beginning of the test and a few difficult items at the end. The easier items will act as positive reinforcers especially in the case of weak students and the more difficult items will provide a challenge to the brighter students and counter the phenomenon known as "ceiling effect".

As an example, let us say that a multiple-choice test has been administered to 30 students. A certain item in that test was answered correctly by 20 students. Then the **difficulty level** of the item will be

$$\frac{20}{30} \times 100 = 67\%$$

In this manner, difficulty levels can be obtained for each of the items included in the multiple-choice test.

Difficulty levels can range from 0 (no one passed the item) to 100 (everyone passed the item). Remember difficulty levels can also be expressed as proportions ranging from 0 to 1. The above item having difficulty level of 67% is somewhat easy. If items are meant to assess individual differences, then items having difficulty levels hovering around 50% are preferred. On the other hand, if the objectives of instruction are to teach for mastery of content, then one would be pleased if the items (assuming appropriate items have been written) have difficulty levels in the eighties and nineties.

(ii) Discrimination Level

Discriminating power of an item is defined as the characteristic of the item to differentiate between students possessing different levels of achievement. The level of achievement of students can be determined either through an external or an internal criterion. An external criterion could be teacher rankings or scores from another relevant test. An internal criterion is usually the total test score of which the item in question is a part. The following steps are followed in obtaining discrimination index of an item.

a. Arrange test papers from the highest to the lowest scores (in the case of the total score on the test being the criterion).

b. Form two groups (a top group and a bottom group) on the basis of the ordered scores. If you have a large number of examinees, choose the top one-third or one-fourth, the bottom one-third or one-fourth for further analysis. If you have a small number, just divide them into two equal halves. Remember to have an equal number in both groups. If you have an odd number of examinees, just discard the middle one from further consideration in case of having a small number of examinees.

c. Count the number of examinees who chose each option including the correct option.

d. Record the counts according to the example given below. This example is based on eighteen students, nine in each group.

	Options			
Groups	a	b	c	d*
Top	2	1	0	6
Bottom	3	2	1	3

*Correct Option

e. Compute discrimination index for each item by finding the difference between the numbers in the top and bottom groups for the correct option. Then divide the difference by the number in the top or the bottom group (remember, the number is the same in both groups). The computational formula is:

$$\frac{R_t - R_b}{n}$$

where,

R_t - the number in the top group who chose the correct option.

R_b = the number in the bottom group who chose the correct option.

n = number of examinees in either the top or the bottom group.

For the item data given above, the **discrimination index** is:

$$\frac{6 - 3}{9} = \frac{3}{9} = .33$$

It is possible to repeat the process given in (e) above for each distracter. It could be, however, very tedious. A visual inspection of the numbers in both groups can be of help in evaluating each distracter. A "good" distracter is the one that is more frequently chosen by low-scoring examinees than by high-scoring examinees. The reverse is true for the correct option. If a distracter is chosen by a higher number in the bottom group than in the top group, (e.g., in our example, 6 students in the

bottom group while 3 students in the top group answered the item correctly, then the discrimination index will be -.33), then check the scoring key to see if it is correct. Items having negative discrimination values (when the number of students providing the correct response in the top group is smaller than the corresponding number in the bottom group), should either be replaced or modified after scrutinizing the scoring key. If a distracter is not chosen by anyone or by an equal number in both groups, then consideration should be given to replacing that distracter,[e.g., in our example, if 6 students in the top group while 5 students in the bottom answered the item correctly, (values in the cells for the bottom group will change), then the discrimination index will be .11). Items having low discrimination values (e.g., ranging from 0 to .20) are of little value in providing any useful information. Such items are potential candidates for revision or replacement. All this information is useful only if there do exist individual differences in the ability or skill being measured. As variations among students decreases due to acquisition of knowledge and skills, the discrimination indices will tend to decrease.

Item analysis results should be used only as a guide and not as a sole criterion for selecting items. Before replacing or discarding an item, consideration should be given to matters such as content coverage. When modifying a test, it should be made sure that the nature of the test as determined by the table of specifications is not altered too much. If this happens, the content validity of the test might suffer.

III. TEST RELIABILITY

(i) Introduction

Information is the basic ingredient for all sorts of decision-making. In order for decisions to be accurate, relevant and dependable information is needed. Most decisions made in education require information from tests, examinations, and other systemic observations. Decisions based upon such information will be accurate to the extent to which the information (e.g., marks or scores) yielded by these examinations and tests are reliable and valid. For instance, if a selection committee of a professional college is faced with choosing those applicants whose chances of success at college are quite high, then the selection committee may use examination marks as a criterion for selection. If the examination marks contain a large component of error due to unreliability (e.g., marks derived from poorly-conceived and poorly-conducted examinations) and thus are not reliable and valid as indicators of the candidates' performance, then the use of those marks will lead to erroneous decisions on selection. It is only when adequately reliable and valid testing instruments are used, it is possible then that the readings from these instruments will form a sound basis for making decisions.

Test reliability[1] refers to the dependability or trustworthiness of the test scores. In other words, do the test items yield information which is consistent across time (stability or test-retest reliability), across equivalent samples from the same domain of content (parallel-form reliability), or within a single sample from a specific domain of content (internal-consistency reliability). Another situation in the study of reliability arises when two or more equally knowledgeable individuals observe the same phenomenon (e.g., mark independently answers to a specific essay question) and come to a conclusion (inter-observer/inter-rater/inter-marker reliability).

Let us look at reliability in a different way. Think of a telephone call made between Peshawar and Karachi. The intent of the telephone call is to convey a message through the telephone signal with the hope that will be understood by the receiver. This situation can be expressed in the following form:

Message = Signal + Noise

The quality of the message in terms of being understood is a function of the strength of the signal, that is, less noise on the line. On the other hand, if there is a lot of noise on the line, the message will be less understood. In order to improve the quality of communication by telephone, the noise on the line has to be reduced by eliminating sources of such noise.

Now let us look at a test score along similar lines. A test score (X) of a person may be thought of as consisting of a signal (true part which is unknown) and noise (error part, which is also unknown). The test score X provides an estimate of the true score, T. In other words,

$$X = T \times E$$

The goodness of X as an estimate of T is a function of how large E is; the smaller the E, the better X is, as an estimate of T. In order to obtain good estimates of T's for all students, we have to reduce E's by following good test-development procedures, good test-administration procedures, and good test-scoring procedures. In other words, to make our tests as much as reliable as humanly possible.

In the following section, a brief description of different types of test reliability are presented. These different types are: test-retest, parallel-form, internal consistency, and inter-marker reliability.

---

[1]In order to understand and interpret test reliability usually given in tl orm of a correlation coefficient, some basic knowledge of the concept (orrelation is essential.

(ii) Test-Retest Reliability

If a test is given on one occasion, will the students maintain the same ranks or positions if the test is repeated on a second occasion?

Here we are addressing the issue of the stability of test scores. Put it simply, two measures for the same individuals are obtained by giving the same test twice. The correlation between the two sets of scores yields a test-retest reliability coefficient.

The major problem with this method of estimating reliability is that it does not take into account the errors of measurement which might have occurred due to the content included in the test. Items in a test are a sample from a specific domain of content and subject to sampling errors. Another problem is related to the question of time interval between the two administrations. If the time is too short, memory may play a role in responding to the items on the second administration and produce scores which are alike. On the other hand, if the time is too long, maturation and learning may alter the ranks of the individuals on the second administration of the test and thus affect reliability.

Consider the two sets of data given in Table 1 on computing a correlation coefficient. If the data in set X is considered as scores of the 20 individuals on a test containing 10 items on one occasion and the data in set B are scores of the same individuals on the second occasion (usually a two-week interval between both occasions), then the correlation coefficient of .80 is called a test-retest reliability coefficient (coefficient of stability).

(iii) Parallel-form Reliability

In order to remediate the problem due to content sampling, two samples of items from the same domain of content are obtained under the assumption that these two samples will yield equivalent scores. If the two samples of items (forms) are given to the same students, then the correlation between the two sets of scores is called the parallel-form reliability coefficient.

There are problems with this method also. One, this method does not take into account variation or errors of measurement due to time-sampling. Second, just deriving two samples of items from the same domain of content does not ensure equivalence.

Consider the two sets of data given in Table 1. If the data in set X is considered scores on Form A containing 10 items and the data in set Y is considered scores on Form B containing an equivalent set of 10 items (for the same groups of individuals), then the coefficient of .80 is called a parallel-form reliability coefficient (coefficient of

equivalence).

(iv) Internal Consistency Methods

There are practical problems with the test-retest and parallel form techniques in addition to the problems related to each method. It is often difficult to administer the same test again to the same individuals or to obtain two parallel forms of a test and administer the two forms on two separate occasions. It will be quite helpful if an estimate of the reliability of a test can be obtained from a single test and a single administration. Information from a single administration of a test can be used in two ways to estimate the reliability of the test. One, the test can be split into two halves which may be considered equivalent. This is known as the split-half method. Second, information from each item in terms of its relationships to other items in the test can be used to estimate reliability. This is generally known as the Kuder-Richardson method.

(a) Split-half Method

Under this method, the test is split into two equivalent halves and scored separately. The scores of the two halves are then correlated. This correlation however is an estimate of the reliability of only one-half of the test. In order to obtain the reliability of the complete test, the Spearman-Brown Prophecy formula is used. This formula is:

$$r_c = \frac{2\, r_h}{1 + r_h}$$

where r_c is the reliability of the complete test and r_h is the reliability of the half test. We should be aware of the fact that there are many ways in which a test can be split into two halves. Each split will yield a different reliability coefficient. For practical purposes, if a test contains sufficient number of items and care is taken in arriving at the two splits, the reliability coefficient will be a reasonably good approximation.

The split-half method is sometimes referred to as the odd-even method. If the odd-numbered items are allocated to one half and the even-numbered items are allocated to the other half, then we will have a split on odd-even basis. This odd-even procedure should not be followed blindly because the odd-numbered items in the test may not be equivalent or parallel to the even-numbered items in the test. The goal should always be to obtain two equal splits of the test which could be considered equivalent.

Let us once again consider the correlation obtained in Table 1 from the two sets of data (think of it now as scores of the same individuals as scores on two equal halves). Using the above formula, the reliability coefficient will be .90 (the total test now contains 20 items). We can see from this formula that, in general, longer tests (e.g. lengthened twice as much as we did in this case) will be more reliable than shorter tests. This means that one should not expect a test score to be a reliable indicator of a student performance if it is based upon a test containing a small number of items. In order to obtain test scores which would be reliable, we have to use an adequate number of items in the test. For example, a mathematics test containing 10 items will not be as reliable as a test containing 50 items from the same domain of content.

(b) KR 20 and KR 21 Methods

Kuder and Richardson presented a series of formulae for estimating reliability of a test. Their formula known as KR 20 came to be widely used. It is as follows:

$$KR\ 20\ Reliability = \frac{k}{k-1}\left[1 - \frac{\sum_{k-1}^{k} p_i q_i}{S^2}\right]$$

where k is the number of items in the test, p is the proportion of individuals passing item i, q is the proportion of individuals failing item i($q_i = 1 - p_i$), and S^2 is the variance of the scores on the total test. This formula is applicable in situations if a numerical value of 1 is assigned to a correct response and a value of 0 is assigned to an incorrect or omitted response to each item. KR 20 is based upon the interrelationships of the items in the test. If the items in the test do not intercorrelate significantly, then the reliability coefficient will be smaller compared to the one for which the inter-item correlations are quite high. In other words, it could be expected that items of a more homogenous structure would yield a higher internal consistency reliability coefficient than items of a less homogenous structure.

Consider information in Table 2 for computing a KR 20 (internal consistency) reliability-coefficient. The information in this table has been taken from the administration of a 30-item

Table 2
Example of Computing an Internal Consistency Reliability Coefficient

Item No.	p	q	pq
1 .793	.207	.164	
2 .685	.315	.216	
3 .945	.055	.052	
4 .910	.090	.082	
5 .973	.027	.026	
6 .700	.300	.210	
7 .935	.065	.061	
8 .850	.150	.126	
9 .850	.150	.126	
10.647	.353	.228	
11.565	.435	.246	
12.970	.030	.029	
13.928	.072	.067	
14.865	.135	.117	
15.933	.067	.063	
16.947	.053	.050	
17.965	.035	.034	
18.825	.175	.144	
19.840	.160	.134	
20.480	.520	.250	
21.785	.215	.169	
22.895	.105	.094	
23.973	.027	.026	
24.990	.010	.009	
25.840	.160	.134	
26.855	.145	.124	
27.955	.045	.043	
28.780	.220	.172	
29.933	.007	.007	
30.670	.330	.221	

test to 400 students. What the information in the preceding Table, for example, tells us is that about 79% of the students correctly answered the first item. Item #24 was the easiest while item #20 was the most difficult for this group of students.

If we refer back to the formula (KR20), we require Σ pq (q = 1-p), and the value of variance (S^2). For our test-score data, Σ pq is 3.424 and S^2 is 14.484. Therefore, the KR20 estimate of the reliability is

$$\frac{30}{29}\left(1 - \frac{3.424}{14.484}\right)$$

$$1.0345\ (1 - .2364) \qquad 1.0345\ (.7636) = .$$

We can say that the internal consistency reliability is in the moderate to high range. The value (reliability coefficient = .79). The value (reliability coefficient = .79) is the same yielded by the analysis done by using the computer (see last page of Appendix B). The KR 20 reliability coefficient is also known as coefficient Alpha.

There is an approximation to KR 20, known as KR 21 which is easier to use. The KR 21 requires the knowledge of the mean and the variance of the test scores. The formula for KR 21 is given below.

$$KR\ 21 = \frac{k}{k-1}\left[1 - \frac{\bar{X}\ (k - \bar{X})}{k\ S^2}\right]$$

In our example, $\bar{X} = 25.34$, $S^2 = 14.484$, $k = 30$

$$\therefore KR\ 20 = \frac{30}{29}\ [1 - \frac{25.34\ (30 - 25.34)}{30 \times 14.484}]$$

$$= 1.035\ (1 - \frac{118.0844}{434.52})$$

$$= 1.035\ (1 - .2718)$$

$$= 1.034\ (.7282) = .75$$

Note the KR 21 estimate of .75 is less than KR 20 estimate of .79. It will always be so, because certain assumptions have been made for the test score data in the case of KR 20 which were not made in the case of KR 21. Therefore, if you wish a more accurate estimate, use KR 20 instead of KR 21. The computer yields KR 20 estimate of the reliability known as alpha.

(v) Reliability of Essay Questions (Inter-marker Reliability)

The methods presented above apply only in situations in which the items can be scored objectively and dichotomously as right or wrong. That is, the scoring procedure is not a source of variation in the test scores. However, scoring is one of the major sources of variation in essay-type tests. In order to study the reliability of scoring of essay-type questions, two or more independent examiners are employed

and then the correlation between, or among, the multiple scores assigned to the response to an essay question provides an estimate of the reliability of the examiner's judgements. This will be the reliability estimate for a single question. In essence, such reliability coefficient indicates the extent to which two or more examiners agree in judging the same response. This procedure can be repeated for all questions and by averaging the reliability coefficients, one will obtain the overall reliability of the essay test.

Consider the two data sets in the earlier example on correlation as marks assigned to an essay response of 20 students. The correlation between the two sets of marks indicates the inter-marker reliability. This is the extent to which both markers agree on the relative status of the individuals in the group on their responses to that specific essay question.

(vi) Interpreting Reliability

Reliability coefficients can theoretically range from 0 (complete lack of reliability) to 1 (perfect reliability). The two extremes are, however, almost never encountered in practical situations. (A correlation coefficient which is the basis of a reliability coefficient, ranges from -1 to +1). Ordinarily, one should expect that measures of achievement should yield reliability coefficients in the range of .80's and .90's.

Another way of interpreting a reliability coefficient is in terms of the errors of measurement. If the reliability of a test is, say, .91 then the magnitude of error variance is .09 or nine percent. If we know the standard deviation of the test scores, then we can translate the error variance into the Standard Error of Measurement (SEM) which facilitates the interpretation of an observed test score. Let us say that the standard deviation of the test with reliability .91 is 10. Then

$$\text{SEM} = SD\sqrt{1 - r}$$

$$10\sqrt{1 - .91}$$

$$= 10\sqrt{.09}$$

$$= 10 \times .3 = 3$$

An observed score X is an estimate of the true score (T). Knowing that the SEM is 3.0, by using the properties of normal distribution (see Appendix F), we can say that the chances are two out of three that the true score will fall in the range X $\pm$ 3. If we wish to be more accurate in our statement about the true score, we can extend the range by saying that the chances are 19 out of 20 that the true score will fall in range

of X ± 6. The value 3 in the first instance and the value 6 in the second instance are known as margin of error in estimating the true score with 68% confidence and 95% confidence respectively.

IV. TEST VALIDITY

Each test is designed and administered to serve a specific purpose such as the extent to which students have achieved a certain objective, e.g., multiplying 2 digit numbers with no carry-over. The purpose for which a test is given may not be as simple and specific as the example given above. Another way (somewhat complex) is to define validity as the extent to which certain inferences can be drawn accurately from test scores. In other words, if we give a test in algebra (assuming that the limits of the domain of content are well-defined),then the test score of an individual should enable us to make an accurate inference regarding the knowledge of that individual in the content of algebra under consideration. You may have come across situations where students complain about the unfairness of the test, that is, the test being out of the course. What they are questioning is the validity of the test in that accurate inferences cannot be made on the basis of their test performance? For a test to be valid, it has to be reliable. How can we make accurate inferences if the test is not yielding consistent information. However, a test can be reliable but not valid. That is, we can have a reliable weighing scale but it will not be a valid measure of height.

There are different types of validity (content, predictive, concurrent, construct, etc.) depending upon the purpose for which the test is developed. We will concentrate on content validity as this type of validity is most crucial for achievement-type tests.

As mentioned earlier, the purpose of a test may be to make inferences about the performance (competence) of an individual in a certain domain of content (algebra). The test as a whole serves as a sample if the items in the test are drawn from the domain of content which is clearly and fully defined. Then content validity of the test indicates the extent to which the test represents that specific domain of content.

In determining content validity, we must first define the domain of content which is of interest to us. In doing this, we should consider both the subject matter and the mental skills (e.g., recall,. . . application,. . . synthesis. . .) which we see as desirable outcomes of instruction and which have been emphasized in teaching (the earlier presentations on specifications and blueprint are relevant here to ensuring content validity of a test). There has been some controversy on the level of specificity to which we should go in defining subject matter and mental processes. This task is easier in some domains but difficult in other domains (e.g., math vs English). A reasonable requirement will be that the test developer attempts to specify the domain and mental processes with considerable detail to enable another individual to examine the representativeness of the test items as a sample from that specific

domain of content and the type of mental skills as outcomes of the teaching-learning process.

Content validity is determined in a subjective manner, generally from the examination of the test items. Other types such as predictive and concurrent validities can be established empirically through obtaining correlations between various data sets. If one is too ambitious, one may ask two knowledgeable judges (experts) to make judgment (content and process separately) on each item (using a rating scale) and then compute the interjudge agreements in terms of correlations (one for content and the other for process). In order to determine content validity, the test developer has to provide information on his/her sampling procedures used in sampling the domain of content (e.g., a content by process matrix of specifications). The technique of internal consistency reliability also tells us whether the test items are homogenous in terms of content. This, however, does not tell us that the test items are a representative sample from the specified domain of content.

In a nutshell, detailed definition of the domain of content and mental skills, representative samples from the domain of content and inventory of mental skills in the form of a content by process matrix, and appropriate procedures for developing test items to specifications will ensure an acceptable level of content validity of a test.

V. INTERPRETING AND USING TEST RESULTS

(i) Introduction

Now that the test has been scored and analyzed, it is time to examine the results and to see what they mean. In other words, what kind of information has been provided by the results and how we can use it in improving the quality of instruction and in monitoring the system.

(ii) Item-level Results

Refer to Appendix C in which complete analysis of test data based upon 60 items and 1685 students is given. The first column just serializes the items.
Values biserials and point-biserials provide information on the discriminating power of each item (see notes on item analysis). A value in these columns tells us the extent to which the item has discriminated between high-scoring and low-scoring students; an indication of item validity. These are correlations (not exactly the type we have discussed earlier) indicating the extent of agreement between the item-scores and the total scores of the examinees.

Why two values? Well, one can obtain two types of correlations between an item-score and the total score depending upon whether one wishes to make certain

assumptions about the item-score. Values in the column label "biser" (biserial) have been obtained by assuming that the ability/skill measured by the item is continuous with an underlying normal distribution. The point-biserial correlations (given in column labelled "point biser") have been obtained where such assumptions have not been made. Remember, a point-biserial correlation will always be less than a biserial correlation. Now, you may be asking which one to use for interpreting the discriminating power of the items. Well, there is no clear agreement on this issue. I would suggest that you use the biserial, because an item in the test measures an ability which can be measured on a continuum with an underlying normal distribution. This emphasis makes sense when there do exist individual differences on the ability being tested.

If concepts have been taught which everyone in the group would have mastered, then the analysis should be limited to the interpretation of proportions endorsing each alternative including the correct response. The examination of proportions for distractors provides useful diagnostic information. If there are students who have opted for these distractors, then such information tells us the kind of mistakes students make. Attempts can be made through instruction to rectify these mistakes. Having said this, it is quite rare that we will encounter situations in an educational system that almost all our students would have mastered a certain skill. There will always be individual differences reflected in the test performance of students no matter how hard one tries to teach the skill. See results for NWFP in Appendix E for examples.

(iii) Test-Level Results

The computer program that is used for scoring and item analysis also provides summary statistics. Refer to the last page of Appendix E. The test contained 60 items and was given to 1685 students. It was an achievement test and scored by assigning 0 to an incorrect response and 1 to a correct response. Therefore, the total score of a student was just the simple sum of the correct response. The average (mean) is 31.84 which is about 53% of the total test score. The standard deviation is 11.06 which indicates that there were variations within the student group (the scores ranged from 0 to 58; see minimum and maximum values). Variance is another measure of variability, it is just the square of the standard deviation. Skewness and kurtosis are other characteristics of the distribution (skewness is of interest to us). The skewness of the distribution is almost zero. This indicates that the students, in general, have performed at the average level.

The alpha coefficient of .91 indicates the internal consistency reliability of the test. The test is fairly consistent, that is, the items generally represent the same domain of content and performance of the group on the items is correlated. The SEM (Standard Error of Measurement) given here (you need information on standard deviation and reliability for computing this) is used to make confidence statements about the

likelihood of the score as an estimate of the true score. The mean p is the average difficulty value. It is obtained by summing the difficulty levels of the individual items and then dividing the sum by the number of items. If you multiply this value by the number of items, you will obtain the arithmetic mean. This value (mean p) tells us that the performance in the test was at the average level. The mean point-biserial is the average of the correlations of items with total scores when item scores are considered discrete. The mean biserial is the average correlation between the items and the total scores when item scores are considered continuous with an underlying normal distribution. These indices tell us that, items in this test possess adequate validity.

In Pakistan, there is a tradition of school-based and centralized examinations at every step of the educational system but the results from these tests and examinations are never analyzed. If tests and examinations are to be useful for purposes other than assigning a grade or making promotion and selection decisions, then the data yielded by tests and examinations have to be analyzed assuming that we have used adequately reliable and valid tests (analyses are also essential for improving the tests and examinations for subsequent use). In other words, there has to be in place a systematic program of psychometric and evaluative research to guide us in developing quality assessment instruments and in making decisions (academic, policy, other types), which are based upon reliable and valid information which hopefully should lead to significant improvements in the overall quality of education.

APPENDIX D:

GLOSSARY OF TERMS

Achievement Test: *A test designed to measure the amount of knowledge and/or skill a person has acquired, usually as a result of planned instruction.*

Average: *A general term used to indicate measures of central tendency such as mean, median, and mode.*

Continuous variable: *A variable that can assume any value within a specified range. Distance, weight, height, time, etc. are examples of continuous variables. Test scores are considered continuous variables.*

Content Validity: *The content validity of a test is indicated by the extent to which the items in the test can be considered a representative sample from a specified domain of content. This type of validity is usually established through content analysis of the items. The determination of content validity is quite important for achievement tests.*

Correlation Coefficient: *An index of association or agreement between two variables, that is, the tendency for values of one variable to change systematically with changes in values of another variable. The existence of a relationship between two variables does not necessarily means causation. A correlation coefficient is an absolute measure of relationship, that is, correlations between the same variables from different studies are comparable. A correlation coefficient can assume values from -1.00 (perfect negative correlation) to +1.00 (perfect positive correlation). A correlation around zero indicates the absence of relationship between the two variables. Different types of correlations can be computed depending upon the nature of the data. The most commonly used coefficient is the Pearson product-moment correlation (r) requiring both variables to be continuous with an underlying normal distribution. The biserial correlation (bisr) assumes that one variable is continuous with an underlying normal distribution (however, the values of this variable are reported categorically) and the other variable is continuous with an underlying normal distribution. The point-biserial correlation (pbisr) assumes that one variable is categorical (discrete) and the other variable is continuous with an underlying normal distribution.*

Criterion-Referenced Test: *A test designed to provide a measure of performance that is interpretable in terms of a clearly defined and delimited domain of learning tasks. Other terms that are less often used but have meanings similar to criterion-referenced: content referenced, domain referenced, and universe referenced.*

Difficulty Index: *An indication of a test item's difficulty generally expressed as the percentage or proportion of individuals who answered the item correctly. The larger the magnitude of the difficulty index, the easier the item. The smaller the difficulty index, the more difficult the item.*

Discrimination Index: *This is an indication of the validity of an item. It is determined by the extent to which an item differentiates between individuals with varying level of relevant ability or skill.*

Discrete variable: *A discrete variable can assume specific values within a specified range. Gender, marital status, political or religious affiliation are examples of discrete variables. Sometimes, a variable is measured on a continuous scale, but treated discretely for the purpose of analysis. Dividing students into high, medium, and low groups on the basis of test scores is an example of artificial categorization.*

Distracter: *An incorrect option for a multiple-choice item.*

Error: *A generic term used to indicate influences which operate in a test and testing situation adversely affecting the reliability and validity of inferences based upon performance on tests and tasks. There are generally two types of error: systematic and random. Systematic error is usually the same for every individual, e.g., giving an extra score to every individual, allowing extra three minutes for completing the test, etc. Random error on the other hand is due to chance,e.g., someone having a headache on the day of the test, a test item is interpreted differently by different individuals, errors due to sampling from the domain of content, etc. Systematic error can be controlled through careful test development and test administration while random error cannot be controlled, but it can be estimated.*

Essay Test: *A test containing a few long and/or short-answer questions to which the examinees provide responses usually in written form. The scoring of an essay test is subjective, that is, rarely, will two or more scorers arrive at the same score for a response of a person to a specific essay question.*

Intermarker Reliability: *The term is used to indicate the extent to which two or more markers would agree upon the marks assigned to responses to a single essay-type question for a group of individuals. It is usually given in the form a correlation coefficient. The terms interexaminer, interscorer, and interrater are also used to indicate this type of reliability.*

Internal Consistency: *A term referring to any of several methods of establishing reliability of a test through the analysis of items representative of a single domain of content.*

Item: *A structured problem or task on a test. The term is commonly used in the context of an objective test.*

Item Analysis: *The process of analyzing the responses of a group of examinees to a test item to determine the difficulty value and discrimination value of the test item.*

Kuder-Richardson Formula: *Any of several formulae developed by Kuder and Richardson for the obtaining internal consistency estimates of test reliability.*

Mean: *It is also called an average. The most commonly used measure of central tendency. It is obtained by summing the scores and dividing the sum by the number of examinees in the group.*

Normal Distribution (curve): *It is also known as normal probability curve. It is a mathematical model which represents the distribution expected for a large number of random observations. Although a normal distribution can never be obtained for a sample of observations,it is quite reasonable to assume a normal distribution for the population from which the sample of observations is drawn. The normal distribution has known properties. For instance, approximately 34% of the area under the normal curve lies between the mean and one standard deviation to the right of the mean. Similarly, the same amount of area lies between the mean and one standard deviation to the left of the mean because the normal curve is symmetrical.*

Norm-Referenced Test: *A test designed to provide a measure of performance that is interpretable in terms of an individual's relative standing in some known group.*

Objective-Referenced Test. *A test designed to provide a measure of performance that is interpretable in terms of a specific instructional objective.*

Objective Test: *A test for which there is complete agreement on the scoring procedures. The same score will be obtained by an examinee no matter who scores the test.*

Observed Score: *The term refers to the score obtained by scoring the test, e.g., summing the number of correct responses to items in an objective test. It is called a raw score.*

Parallel-form Reliability: *A reliability coefficient obtained by correlating scores for a group of examinees on two tests sampling the same domain of*

content. The terms alternate-form or equivalent-form are also used to refer to this type of reliability.

Percentile Rank: *The percent of scores falling below a given score in a frequency distribution of scores.*

Population: *A total group of interest or concern so defined. Usually, a subset (randomly drawn) from a population is studied and inferences are made about the population on the basis of the results obtained for the subset.*

Random Sample: *A sample drawn from a population in such a way that every element of the population stands an equal chance to be included in the sample.*

Range: *It is calculated by obtaining the difference between the highest and lowest scores on a test.*

Reliability: *A term referring to the consistency with which a test measures what it is intended to measure. The reliability of a test is generally reported in the form of a correlation coefficient. The different types are: test-retest, parallel-form, internal consistency, intermarker (for essay-type tests), etc.*

Scoring Key: *The list of correct answers for the items in a test.*

Skewness: *The skewness gives an indication of the shape of the score distribution. A negative skewness indicates that there is a relative abundance of scores at the high end of the scale distribution. A positive skewness means that there is a relative abundance of scores at the low end of the distribution. A skewness of zero means that the scores are more symmetrically distributed about the mean.*

Spearman-Brown Prophecy Formula: *A formula developed to estimate the reliability of a test if it is lengthened but keeping the other factors constant (e.g., nature of content, format of items, difficulty level of items, etc.). It is most commonly used in "correcting" the split-half reliability coefficient.*

Split-half Reliability: *An estimate of test reliability based on the correlation between scores on two equivalent halves of the test.*

Standard Deviation: *A measure of variability or spread of a distribution generally preferred over other measures of variability because of its sound mathematical basis and usefulness in interpreting test scores. The squared value of the standard deviation is called variance.*

Standard Error of Measurement: *An estimate of the standard deviation of the distribution of scores that would be obtained if it was possible to test a person again and again on the same test assuming no intervening experiences relevant to the test.*

Stem: *That part of a multiple-choice item which contains the problem or task. The stem can be presented as an incomplete statement or as a question.*

Test-retest Reliability: *A reliability coefficient obtained by correlating scores of a group of examinees on two administrations (generally with a two-week interval) of the same test with no intervening experiences relevant to the test.*

True Score: *A theoretical score never achievable in practice; usually defined as the average of the scores if it were possible to administer the same test to a person infinite number of times assuming no intervening experiences relevant to the test. The observed score of a person on a single administration of the test is an estimate of the true score.*

Validity: *It refers to the extent to which a test serves the purpose for which it was designed. There are many types; e.g., content, concurrent, predictive, construct, etc.*

Variable: *It refers to a characteristic or trait which can assume different values within a specified range.*

APPENDIX E: Item and Test Statistics

Item No.	Alt.	Item Statistics: Prop. Endorsing	Biser.	Point Biser.	Key
1	1	0.847	0.500	0.328	*
	2	0.056	-0.367	-0.181	
	3	0.049	-0.313	-0.147	
	4	0.033	-0.388	-0.160	
	Other	0.014	-0.440	-0.135	
2	1	0.082	-0.234	-0.130	
	2	0.463	0.110	0.087	
	3	0.278	-0.311	-0.233	
	4	0.154	0.470	0.309	*
	Other	0.023	-0.273	-0.099	
3	1	0.245	-0.166	-0.121	
	2	0.111	-0.213	-0.128	
	3	0.515	0.390	0.311	*
	4	0.090	-0.198	-0.112	
	Other	0.040	-0.362	-0.159	
4	1	0.097	-0.165	-0.096	
	2	0.205	-0.434	-0.305	
	3	0.049	-0.310	-0.146	
	4	0.625	0.544	0.426	*
	Other	0.024	-0.418	-0.154	
5	1	0.149	-0.228	-0.149	
	2	0.562	0.540	0.429	*
	3	0.096	-0.263	-0.152	
	4	0.082	-0.314	-0.174	
	Other	0.111	-0.356	-0.215	
6	1	0.185	-0.065	-0.045	
	2	0.709	0.346	0.262	*
	3	0.043	-0.491	-0.222	
	4	0.015	-0.309	-0.097	
	Other	0.048	-0.443	-0.207	
7	1	0.023	-0.226	-0.082	
	2	0.383	-0.234	-0.184	
	3	0.519	0.414	0.331	*
	4	0.040	-0.290	-0.128	
	Other	0.036	-0.493	-0.209	

Alternative Statistics

Seq. No.	Alt.	Prop. Endorsing	Biser.	Point Biser.	Key
8	1	0.154	0.011	0.007	
	2	0.272	0.342	0.255	*
	3	0.223	-0.125	-0.090	
	4	0.228	-0.064	-0.046	
	Other	0.122	-0.294	-0.182	
9	1	0.082	-0.247	-0.136	
	2	0.037	-0.235	-0.101	
	3	0.072	-0.289	-0.154	
	4	0.706	0.455	0.344	*
	Other	0.102	-0.339	-0.199	
10	1	0.231	-0.214	-0.155	
	2	0.048	-0.216	-0.101	
	3	0.382	0.569	0.446	*
	4	0.246	-0.242	-0.177	
	Other	0.093	-0.323	-0.185	
11	1	0.364	0.001	0.001	
	2	0.122	-0.262	-0.162	
	3	0.116	-0.221	-0.135	
	4	0.303	0.474	0.360	*
	Other	0.095	-0.411	-0.237	
12	1	0.323	-0.157	-0.121	
	2	0.027	-0.373	-0.144	
	3	0.023	-0.356	-0.129	
	4	0.576	0.397	0.314	*
	Other	0.052	-0.532	-0.254	
13	1	0.779	0.548	0.392	*
	2	0.103	-0.381	-0.225	
	3	0.031	-0.349	-0.140	
	4	0.045	-0.319	-0.147	
	Other	0.040	-0.449	-0.198	
14	1	0.021	-0.328	-0.115	
	2	0.906	0.562	0.323	*
	3	0.022	-0.473	-0.169	
	4	0.018	-0.383	-0.129	
	Other	0.033	-0.488	-0.201	

Alternative Statistics

Seq. No.	Alt.	Prop. Endorsing	Biser.	Point Biser.	Kev
15	1	0.064	-0.241	-0.123	
	2	0.059	-0.282	-0.141	
	3	0.112	-0.258	-0.156	
	4	0.709	0.482	0.364	*
	Other	0.056	-0.465	-0.229	
16	1	0.557	0.415	0.330	*
	2	0.079	-0.085	-0.046	
	3	0.042	-0.186	-0.083	
	4	0.284	-0.277	-0.208	
	Other	0.039	-0.490	-0.213	
17	1	0.050	-0.113	-0.053	
	2	0.034	-0.333	-0.139	
	3	0.761	0.316	0.230	*
	4	0.066	-0.145	-0.075	
	Other	0.087	-0.265	-0.149	
18	1	0.062	-0.334	-0.170	
	2	0.633	0.643	0.502	*
	3	0.080	-0.375	-0.205	
	4	0.041	-0.322	-0.143	
	Other	0.185	-0.440	-0.303	
19	1	0.027	-0.372	-0.144	
	2	0.625	0.613	0.480	*
	3	0.165	-0.331	-0.221	
	4	0.127	-0.303	-0.190	
	Other	0.056	-0.567	-0.278	
20	1	0.037	-0.351	-0.151	
	2	0.890	0.652	0.392	*
	3	0.015	-0.535	-0.166	
	4	0.031	-0.405	-0.164	
	Other	0.027	-0.736	-0.282	
21	1	0.651	0.169	0.131	
	2	0.099	-0.423	-0.247	
	3	0.035	-0.456	-0.192	
	4	0.195	0.266	0.185	*
	Other	0.021	-0.542	-0.190	

Alternative Statistics

Seq. No.	Alt.	Prop. Endorsing	Biser.	Point Biser.	Key
22	1	0.040	-0.268	-0.118	
	2	0.059	-0.226	-0.113	
	3	0.497	0.732	0.584	*
	4	0.364	-0.548	-0.428	
	Other	0.040	-0.420	-0.185	
23	1	0.417	0.763	0.604	*
	2	0.397	-0.600	-0.473	
	3	0.025	-0.243	-0.091	
	4	0.098	-0.030	-0.018	
	Other	0.063	-0.379	-0.193	
24	1	0.082	-0.291	-0.161	
	2	0.056	-0.313	-0.154	
	3	0.078	-0.215	-0.117	
	4	0.675	0.568	0.436	*
	Other	0.107	-0.503	-0.300	
25	1	0.874	0.612	0.382	*
	2	0.039	-0.470	-0.204	
	3	0.046	-0.367	-0.169	
	4	0.024	-0.399	-0.148	
	Other	0.017	-0.688	-0.225	
26	1	0.011	-0.338	-0.093	
	2	0.187	-0.329	-0.227	
	3	0.495	0.808	0.644	*
	4	0.275	-0.559	-0.418	
	Other	0.032	-0.523	-0.213	
27	1	0.117	-0.290	-0.177	
	2	0.499	0.621	0.496	*
	3	0.123	-0.395	-0.245	
	4	0.109	-0.213	-0.128	
	Other	0.153	-0.301	-0.197	
28	1	0.722	0.592	0.444	*
	2	0.019	-0.431	-0.146	
	3	0.012	-0.338	-0.096	
	4	0.221	-0.448	-0.321	
	Other	0.026	-0.580	-0.220	

Alternative Statistics

Seq. No.	Alt.	Prop. Endorsing	Biser.	Point Biser.	Key
29	1	0.496	0.649	0.518	*
	2	0.254	-0.348	-0.256	
	3	0.121	-0.298	-0.184	
	4	0.062	-0.242	-0.123	
	Other	0.066	-0.448	-0.232	
30	1	0.141	-0.140	-0.090	
	2	0.502	0.662	0.528	*
	3	0.289	-0.475	-0.358	
	4	0.020	-0.274	-0.094	
	Other	0.048	-0.575	-0.269	
31	1	0.376	0.772	0.605	*
	2	0.310	-0.398	-0.303	
	3	0.093	-0.264	-0.151	
	4	0.099	-0.207	-0.121	
	Other	0.121	-0.360	-0.222	
32	1	0.589	0.096	0.076	
	2	0.062	-0.338	-0.171	
	3	0.206	0.355	0.250	*
	4	0.090	-0.270	-0.153	
	Other	0.053	-0.492	-0.238	
33	1	0.042	-0.318	-0.142	
	2	0.500	0.485	0.387	*
	3	0.142	-0.172	-0.111	
	4	0.148	-0.317	-0.206	
	Other	0.169	-0.211	-0.142	
34	1	0.487	0.718	0.573	*
	2	0.185	-0.295	-0.203	
	3	0.096	-0.375	-0.217	
	4	0.077	-0.277	-0.150	
	Other	0.155	-0.435	-0.286	
35	1	0.218	-0.124	-0.088	
	2	0.079	-0.320	-0.175	
	3	0.518	0.442	0.353	*
	4	0.126	-0.173	-0.108	
	Other	0.060	-0.479	-0.240	

Seq. No.	Alt.	Alternative Statistics: Prop. Endorsing	Biser.	Point Biser.	Key
36	1	0.285	-0.486	-0.365	
	2	0.630	0.650	0.508	*
	3	0.034	-0.262	-0.110	
	4	0.021	-0.413	-0.146	
	Other	0.030	-0.585	-0.232	
37	1	0.491	0.588	0.469	*
	2	0.064	-0.342	-0.175	
	3	0.065	-0.065	-0.033	
	4	0.331	-0.336	-0.259	
	Other	0.050	-0.600	-0.284	
38	1	0.550	0.785	0.625	*
	2	0.167	-0.511	-0.343	
	3	0.062	-0.097	-0.049	
	4	0.174	-0.450	-0.305	
	Other	0.047	-0.565	-0.262	
39	1	0.153	-0.190	-0.125	
	2	0.073	-0.258	-0.138	
	3	0.653	0.504	0.391	*
	4	0.045	-0.191	-0.087	
	Other	0.077	-0.607	-0.328	
40	1	0.142	-0.200	-0.129	
	2	0.489	0.097	0.078	
	3	0.109	-0.029	-0.017	
	4	0.170	0.363	0.245	*
	Other	0.090	-0.492	-0.280	
41	1	0.057	-0.127	-0.063	
	2	0.024	-0.337	-0.125	
	3	0.589	0.746	0.590	*
	4	0.301	-0.628	-0.477	
	Other	0.029	-0.570	-0.225	
42	1	0.096	-0.118	-0.068	
	2	0.281	-0.313	-0.235	
	3	0.482	0.604	0.482	*
	4	0.099	-0.346	-0.202	
	Other	0.043	-0.598	-0.270	

Seq. No.	Alt.	Alternative Statistics: Prop. Endorsing	Biser.	Point Biser.	Key
43	1	0.050	-0.224	-0.106	
	2	0.078	-0.337	-0.183	
	3	0.763	0.578	0.419	*
	4	0.050	-0.374	-0.177	
	Other	0.059	-0.576	-0.287	
44	1	0.077	-0.332	-0.180	
	2	0.074	-0.269	-0.144	
	3	0.311	-0.300	-0.229	
	4	0.480	0.648	0.517	*
	Other	0.059	-0.569	-0.284	
45	1	0.285	-0.161	-0.121	
	2	0.145	-0.336	-0.218	
	3	0.359	0.577	0.449	*
	4	0.156	-0.097	-0.064	
	Other	0.054	-0.557	-0.270	
46	1	0.635	0.618	0.482	*
	2	0.147	-0.347	-0.225	
	3	0.096	-0.305	-0.176	
	4	0.053	-0.256	-0.124	
	Other	0.069	-0.547	-0.287	
47	1	0.087	-0.142	-0.080	
	2	0.174	-0.074	-0.050	
	3	0.427	-0.025	-0.020	
	4	0.229	0.427	0.307	*
	Other	0.083	-0.508	-0.282	
48	1	0.187	0.006	0.004	
	2	0.077	-0.166	-0.090	
	3	0.351	0.308	0.239	*
	4	0.211	-0.118	-0.084	
	Other	0.175	-0.223	-0.152	
49	1	0.049	-0.135	-0.064	
	2	0.057	-0.345	-0.170	
	3	0.767	0.495	0.358	*
	4	0.074	-0.369	-0.198	
	Other	0.053	-0.428	-0.207	

Alternative Statistics

Seq. No.	Alt.	Prop. Endorsing	Biser.	Point Biser.	Key
50	1	0.121	-0.325	-0.200	
	2	0.154	-0.277	-0.182	
	3	0.534	0.630	0.502	*
	4	0.092	-0.224	-0.128	
	Other	0.099	-0.472	-0.276	
51	1	0.354	0.484	0.376	*
	2	0.135	-0.017	-0.011	
	3	0.230	-0.286	-0.206	
	4	0.198	-0.072	-0.050	
	Other	0.082	-0.455	-0.252	
52	1	0.188	0.219	0.151	
	2	0.539	-0.143	-0.114	
	3	0.091	0.329	0.188	*
	4	0.077	-0.015	-0.008	
	Other	0.104	-0.298	-0.177	
53	1	0.284	0.157	0.118	*
	2	0.303	0.036	0.027	
	3	0.197	-0.250	-0.174	
	4	0.170	0.168	0.113	
	Other	0.046	-0.408	-0.187	
54	1	0.227	0.379	0.273	*
	2	0.335	-0.101	-0.078	
	3	0.261	-0.129	-0.096	
	4	0.116	0.089	0.054	
	Other	0.061	-0.440	-0.222	
55	1	0.178	-0.152	-0.104	
	2	0.464	0.364	0.290	*
	3	0.149	-0.232	-0.151	
	4	0.083	0.123	0.068	
	Other	0.126	-0.339	-0.211	
56	1	0.276	-0.167	-0.125	
	2	0.303	0.464	0.352	*
	3	0.135	-0.011	-0.007	
	4	0.171	-0.132	-0.089	

Alternative Statistics

Seq. No.	Alt.	Prop. Endorsing	Biser.	Point Biser.	Key
57	1	0.082	-0.233	-0.129	
	2	0.052	-0.295	-0.142	
	3	0.763	0.530	0.385	*
	4	0.017	-0.303	-0.098	
	Other	0.086	-0.538	-0.301	
58	1	0.017	-0.334	-0.108	
	2	0.850	0.524	0.342	*
	3	0.032	-0.188	-0.077	
	4	0.031	-0.350	-0.142	
	Other	0.070	-0.523	-0.275	
59	1	0.084	-0.326	-0.181	
	2	0.080	-0.238	-0.130	
	3	0.069	-0.270	-0.142	
	4	0.662	0.609	0.471	*
	Other	0.106	-0.554	-0.329	
60	1	0.100	-0.275	-0.161	
	2	0.038	-0.317	-0.137	
	3	0.673	0.548	0.422	*
	4	0.100	-0.232	-0.136	
	Other	0.088	-0.515	-0.290	

Scale Statistics

Scale:	Math
N of Items	60
N of Examinees	1685
Mean	31.837
Variance	122.311
Std. Dev.	11.059
Skew	-0.001
Kurtosis	-0.616
Minimum	0.000
Maximum	58.000
Median	32.000
Alpha	0.913
SEM	3.269
Mean P	0.531
Mean Point Bis	0.399
Mean Biserial	0.527

APPENDIX F:

STEPS IN FIELD TESTING

PREPARATION
ADMINISTRATION
SCORING
USE OF FIELD TEST DATA

Dr. SAR B. KHAN
AED CONSULTANT
WASHINGTON, D.C.

FEBRUARY, 1994

FIELD TESTING

Objective: In this session you will learn how to conduct a field test of your items before they are used in a real situation. You will learn how to assemble and administer a test on a sample basis, and how to revise the items based on field-test results.

After items have been written, they should be critically reviewed by a team of subject-matter and measurement experts and revised on the basis of their criticisms and suggestions. These items, then, must be tried out on a sample of examinees which represents the group of students to whom the final version of the test will be given.

The tryout of the "experimental" versions of the test will help identify problems with test items and/or administrative procedures and will allow the test developers to make improvements before the real administration. Field tests will help:

1. To identify weak or defective items, specifically, items which are unclear, too easy, too difficult, and so forth.

2. To determine the difficulty level of each item and select items with difficulties appropriate to the purpose of the final test.

3. To provide information on how many items in the final test would be sufficient to serve the overall purpose of the final test.

4. To provide information on appropriate time limits for the final test.

5. To identify weaknesses in the mechanics of test administration, in the directions to examiners and examinees, in the method of recording responses, errors in the text of the items, and so forth.

6. To be able to calculate appropriate statistics such as mean, variance, reliability, and validity estimates.

7. To be able to benefit from the suggestions and criticisms of the test administrators in the field in improving test administration procedures.

A very small pre-field test may be worthwhile to identify gross deficiencies in the test. This pre-field test can be conducted with as few as 10 or 12

students. Generally, test developers prefer to administer the pre-tryout themselves, since they can gather useful information from direct observation and interviews with the examinees. Gross deficiencies, ambiguities, omissions in the directions to the examinees and in the sample of items may be discovered and remedied before the formal tryout.

Field Testing: Formal Tryout

Once information from the pre-tryout stage has been used to improve the test items, they have to be tried on a large sample (100 - 400 examinees) to obtain accurate information on the behavior of each item. Sometimes, the results of the first tryout may reveal the desirability of altering some items and of replacing others. In these circumstances, a second tryout may be necessary to determine the adequacy of the revised items.

(i) Assembling the Items

Before items can be tried out, they must, of course, be assembled into tryout units or tryout forms (there may be more than one form if items are part of an initial item bank). It is important that items are presented in an acceptable format such as presenting options vertically rather than horizontally; using letters to indicate options especially for items involving numerical options; not breaking an item (half given on one page and half on the following page); not overcrowding the text of the items which will make reading of items more difficult, etc.. With the advance in printing technology such as the desk-top publishing, several formats are possible.

(ii) Directions

Detailed directions should be prepared for both the examiners and the examinees. Since the care with which the test is administered depends upon the test administrators, it is important that competent examiners are chosen who will conscientiously carry out the directions provided. Ordinarily, the purpose of the testing and the procedures to be followed should be explained to the examiners well in advance during training sessions. The directions supplied to the examiners should be detailed and clear. They should define the responsibilities of administrative staff and include specific instructions regarding time limits, and the disposition or return of the completed and blank test booklets. If sample exercises are included as part of the test, the examiners should be provided with the correct answer to these problems. The necessity of following directions to the letter as much as possible should be emphasized.

Directions to examinees should be given as part of the test. Sample items should be included as part of the directions. This may be especially

necessary in situations where the examinees are not quite familiar with the format in which the items are presented (e.g., multiple-choice vs. essay). The directions should include all the necessary information such as purpose of the test, time limit, how to record responses, whether students should guess, etc.

(iii) Instructions about Guessing

Guessing on objective-type test items such as multiple choice is a concern to test specialists. There should be clear instructions to the students on whether they will be penalized if they guess. This should be clearly stated as part of the directions. There are methods of correcting for guessing; however, it is probably best to encourage students to try to answer all the items -- even if they have to guess. Any improvement in scores due to indiscriminate guessing is minimal in the case of items having four or five options. On the other hand, if a student is able to eliminate some options through intelligent thinking and guesses on the remainder, he/she should not be penalized if he/she provides a wrong answer.

(iv) Time Limits

One of the purposes of the field testing is to determine reasonable time limits for the final test. We would like all the students to be able to complete most of the items in the test. Flexible and generous time limits should be allowed. Examinees may be asked to mark the item they are working on every 15 minutes to give test developers a sense of how students are proceeding. This information can then be analyzed to arrive at a realistic time limit for the final test.

(v) Cheating

Students should be motivated to do well on the test but not allowed to cheat. Cheating, results in unfairness and makes test scores less reliable because individual differences in scores are not due to the learned abilities, but to cheating. As scores on centralized examinations are used as criteria for decisions on admission to professional colleges and institutions of higher learning, and decisions on employment, the effects of cheating in examinations are far-reaching and are detrimental to the moral fabric and healthy functioning of the society. Therefore, cheating and other malpractices have to be stopped at all cost. If this is not possible, then the whole exercise in conducting examinations is a farce and examinations are of little value in providing reliable and valid information for the intended purpose.

(vi) Other Considerations

A test serves its purpose well if the scores derived from the test accurately reflect the performance of the students in the content area of interest. An observed test score can be thought of as consisting of two components: a true component and an error component. The error component can be further classified into systematic and random. There are different types of sources of error which should be controlled if error has to be minimized. Attempts should be made to reduce as much of the systematic error through adequate planning and administration of the test as possible. Sources which could contribute to variation in scores may be physical, psychological, and other situational conditions under which the test is administered. It has to be made sure that students' performance is not affected by inadequate physical conditions such as poor lighting, noise, or other distractions.

(vii) Reproducing Multiple Copies

After the items are assembled into a test, multiple copies of the test have to be reproduced. Whether the test pages have to be photocopied or printed, it is important to make sure that the text is legible and complete (no top or bottom lines missing, pages not missing, no stray marks of any kind, etc.). If the text contains pictures or diagrams, they should be properly drawn and reproduced. It is not uncommon that one finds poorly drawn pictures and diagrams in textbooks. Although the poor quality of pictures and diagrams given in textbooks can be remedied by the teacher through instruction, such opportunity does not exist in a testing situation. If a student fails to provide a correct response to the item because of the quality of the picture or diagram, then this contributes to the unreliability of the test score and invalid inferences regarding the student's performance.

Scoring the Multiple-choice Test

A scoring key (list of correct answers) should be prepared in advance and its accuracy should be checked by one or two persons who are knowledgeable in the field of content being tested. It is suggested that all items should be given equal weight. Traditionally, a correct response is assigned a numerical value of 1 while an incorrect response is assigned a numerical value of 0. The total score is then a simple sum of the correct responses.

The use of a separate answer sheet is not recommended with young children. It may be considered at the class 5 level provided students have been given sufficient practice in the use of a separate answer sheet. The advantage of separate answer sheets is that scoring is very simple and fast. For instance, holes can be punched on the positions of the correct answers.

This sheet can be placed on top of the student's answer sheet and the correct responses counted.

It is necessary to check to make sure that no double responses are given to an item. If there are double entries for one or two items (of which one entry is correct) in the total test, then the student should have the benefit of doubt and these items should be scored as correct. However, if there are more than two entries for an item, then the item should be scored incorrect.

Instructions should be provided about what to do if the examinee changes his/her mind after recording a response (such as "erase your first answer completely). If responses are recorded on the body of the test paper, then the use of a scoring template is not feasible and scoring should be done by visually examining the responses.

In case of large-scale testing, there are scanning machines available which will read specially designed answer sheets and either provide you with a score on the answer sheet or read the responses directly into a computer for further analyses.

Currently, computer programs are available for mainframe and desktop computers for scoring and analyzing test responses. The use of these programs requires the responses of students to each item be entered into the computer. The analysis of the responses by the computer yields invaluable information on how the items are functioning. This information can be used to improve the test items, as well as to provide feedback to teachers and other educators to help improve various aspects of the teaching and learning process.

Using the Field-test Data

The data from the field testing will be useful in preparing the final test. Field reports submitted by the test administrators should be reviewed and appropriate changes made in the test and in administration procedures, based on their suggestions. Data on time should be analyzed to come up with reasonable and fair time limit for the actual test.

The field test data in terms of item responses and total scores should be analyzed statistically. Item statistics will be helpful in revising or replacing the items which did not perform according to expectations for use in the final test. The next section describes basic statistical procedures for analyzing test data and how the results can interpreted for making decisions on the quality of items in the test.

In assembling, administering, and scoring the final test, the procedures described earlier in connection with field testing will be applicable after incorporating necessary changes on the basis of information obtained through field testing. When the final test is administered to large numbers of examinees as part of a regular assessment program, the test score data should be subjected to appropriate statistical analyses to gain further insights into the behavior of the items. Such information will be very useful in bringing about continuous improvements in the assessment program and in the overall quality of the education system.

APPENDIX G:

LEVELS OF MEASUREMENT

	SCALE	CHARACTERISTICS OF THE SCALE	STATISTICAL POSSIBILITIES OF THE SCALE
Non-Interval Scales	Nominal Scale	A scale which measures in terms of names or designation of discrete units or categories.	Can be used for determining the mode, the percentage values, or the Chi-square
	Ordinal Scale	A scale which measures in terms of such values as more or less, larger or smaller, but without specifying the size of the intervals.	Can be used for determining the mode, percentage, Chi-square, median, percentile rank, or rank correlation.
Interval Scales	Interval Scale	A scale which measures in terms of equal intervals or degrees of difference, but whose zero point, or point of beginning is arbitrarily established.	Can be used for determining the mode, the mean, the standard deviation, t-test, the F-test, and the product moment correlation.
	Ratio Scale	A scale which measures in terms of equal intervals and an absolute zero point of origin.	Can be used for determining the geometric mean, the harmonic mean and the percent variation and all other statistical determinations.

APPENDIX H:

THE NEED FOR STANDARDIZATION OF RAW SCORES

SAR B. KHAN
AED CONSULTANT
WASHINGTON, D.C.

FEBRUARY, 1994

Before discussing the need for standardization of raw scores it seems appropriate to provide a brief background to the measurement process in education. This will help the reader in putting the standardization issue in proper perspective.

I. Educational Decision-Making

All persons involved in the educational enterprise (students, parents, teachers, administrators, political leaders, etc) are required to make decisions. These decisions may range from building a school in a community to promoting a child to the next grade. Decisions are necessary when there exists more than one alternative to choose from.

II. Need for Information to Make Decisions

Decisions cannot be made without having accurate and relevant information. In education, information is obtained through needs assessment studies, tests, and examinations, observations (e.g., checklists, rating scales. etc.) survey questionnaires, interview, self-report inventories, etc. Generally speaking, the use of accurate and relevant information in decision-making should greatly reduce the chance of making incorrect decisions.

III. Measurement, Evaluation, and Assessment

Information from educational instruments is generally described in terms of numbers called raw scores. Measurement is defined as the process of assigning a numerical value to a phenomenon of interest according to some specified rule. We engage in evaluation when we attach qualitative meaning to the measurements. When we have information from multiple sources on a certain phenomenon, we can arrive at an overall assessment with regards tot he status of that phenomenon.

IV. Interpretation of Measurements

Educational measurements lack an absolute zero and a well-defined unit of measurement. Therefore, there is a need to interpret these measurements with respect to some frame of measurement. The two frames of reference commonly used are criterion-referenced and norm-referenced. Criterion-referenced interpretation requires he existence of a priori criterion or standard of performance rationally and/or empirically established and linked to a well-defined domain of abilities and skills. Classroom-based mastery testing, training program outcomes, formative testing, etc. are appropriate situations in which to employ criterion-referenced interpretation. In the

criterion-referenced, absolute meaning (such as a passing score of 60%) is attached to the score.

However, when one collects measurements on different school subjects which are not the same in essential characteristics, criterion-referenced interpretation (i.e., applying a uniform standard to each subject) becomes highly questionable. Frequently, it is more desirable to interpret the score of an individual in each subject with reference to a group of which the individual is a member or can be considered to be a member. Scores in different subjects can be interpreted in relation to performance of the reference or norm group in each subject rather than to an absolute value. This is the case of the norm-referenced interpretation.

V. Need for Standardization

When one has scores for students in different school subjects and wished to develop a profile or aggregate the scores, he/she cannot use raw scores because the tests would generally have different means and variances. The raw scores on each test have to be transformed (e.g., linearly) to the same scale. Then, the transformed scores can be added together or used to develop a profile to discern strengths and weaknesses.

A good example of linear transformation is:

$$F^\circ = 9/5C^\circ + 32$$

The above transformation does not change the temperature in any way but the temperature is expressed in terms of a Fahrenheit scale compared to a Centigrade scale. Similarly, a raw score X on a test (X) having a mean ($\overline{X}$) and a standard deviation (S) by using the following formula:

$$Z = \frac{X_i - \overline{X}}{S}$$

The new Z scores will have a mean of zero and a standard deviation of 1.

When the scores on all tests are transformed in the above manner, then each test will have been brought to the same scale, i.e., a mean of zero and a standard deviation of 1. However, it should be remembered that the total score will have of mean of 0 but the standard deviation will not be equal to 1. In order to standardize the total score, each total score should be divided by the standard deviation of the total scores. Once this is done, the resulting total or overall score has now been scaled to a mean of 0 and standard deviation of 1.

If we can assume that the overall score is normally distributed (if plotted, the curve should approximate a bell-shaped curve), there is a direct relationship between Z scores and percentile ranks (a percentile rank corresponds to a point below which there are a certain percentage of scores; e.g., a score corresponding to the 40th percentile rank will have 40% of the scores falling below that score). For instance, a Z-score of +1 corresponds to the 84th percentile.

Usually, it is difficult for teachers and students to understand Z scores because almost half of them (if we have a normal distribution) will be negative. In order to facilitate understanding on the part of teachers and students, the Z score can be transformed to a T score. The T scores have a mean of 50 and a standard deviation of 10. The Z scores can be expressed in terms of T scores by the following formula:

$$T_1 = 10(Z_1) + 50$$

If we have normal distribution, then T scores will have a direct relationship to percentile ranks. For instance, a T score of 40 will correspond to the 16th percentile while a T score of 60 will correspond to the 84th percentile. It may be quite reasonable to assume that the overall performance will be normally distributed. The basis for this assumption is the fact that we are aggregating a number of scores and we would likely have the number of candidates in the thousands. As a final outcome, one can, then, express the overall performance of a candidate on a national examination in terms of a T score and the corresponding percentile rank.

APPENDIX I:
PERCENT SUCCESS-RATE ON MATH ITEMS FOR NWFP

TABLE # 1
PERCENT SUCCESS-RATE ON MATH ITEMS (K=60) BY PROVINCE (N.W.F.P)
PTC STUDENTS

ITEM NO/CONCEPT	*P-VALUE*	*ITEM NO/CONCEPT*	*P-VALUE*
1. Equivalence : Roman Arabic Numerals	85	14. Common Multiples	91
2. Prime Number	15	15. Common Multiples	71
3. Divisibility	51	16. Least Common Multiples (LCM)	56
4. Prime Factorization	62	17. LCM (Word Problem)	76
5. Place Value Whole Number	56	18. Associated Property	63
6. Fractions (Num/Denominator)	71	19. Fractions (Simplification)	62
7. Fractions (Interpretation)	52	20. Mixed Fractions(simplification)	89
8. Proper/Improper Fractions	27	21. Fractions (Addition)	19*
9. Composite Numbers	71	22. Fractions (Addition)	50
10. Common Divisors	38	23. Fractions (Addition)	42
11. Equivalent Fractions	30	24. Fractions(Mixed + whole No)	67
12. Fractions (Interpretation)	58	25. Fractions (Subtraction)	87
13. Greatest Common Divisor	78	26. Fractions (Subtraction)	49

TABLE # 1 (Contd)

PERCENT SUCCESS-RATE ON MATH ITEMS (K=60) BY PROVINCE (N.W.F.P) PTC STUDENTS

ITEM NO/CONCEPT	*P-VALUE*	*ITEM NO/CONCEPT*	*P-VALUE*
27. Mixed Fractions (Subtraction)	50	40. Decimal Fractions (Place Value)	17
28. Fractions (Multiplication)	72	41. Decimal Fractions (Subtraction)	59
29. Mixed Fractions (Multiplication)	50	42. Decimal Fractions(Multiplication)	49
30. Fractions (Division)	50	43. Decimal Fractions(Multiplication)	76
31. Mixed fractions (Division)	38	44. Decimal Fractions(Multiplication)	48
32. Fractions (Word Problem) *	21	45. Decimal Fractions (Division)	36
33. Fractions (Mixed Operations)	50	46. Decimal Fractions (Division)	63
34. Fractions (Mixed Operations)	49	47. Decimal Fractions (Division)	23
35. Fractions (Order)	52	48. Decimal Fractions (Mixed opera)	35
36. Decimal Fractions (Addition)	63	49. Unitary Method	77
37. Conversion to Decimal Fractions	49	50. Average	53
38. Conversion to Decimal Fractions	55	51. Rays and Angles	35
39. Conversion to Mixed Fractions	65	52. Area (Triangle)	9

TABLE # 1 (Contd)

PERCENT SUCCESS-RATE ON MATH ITEMS (K=60) BY PROVINCE (N.W.F.P) PTC STUDENTS

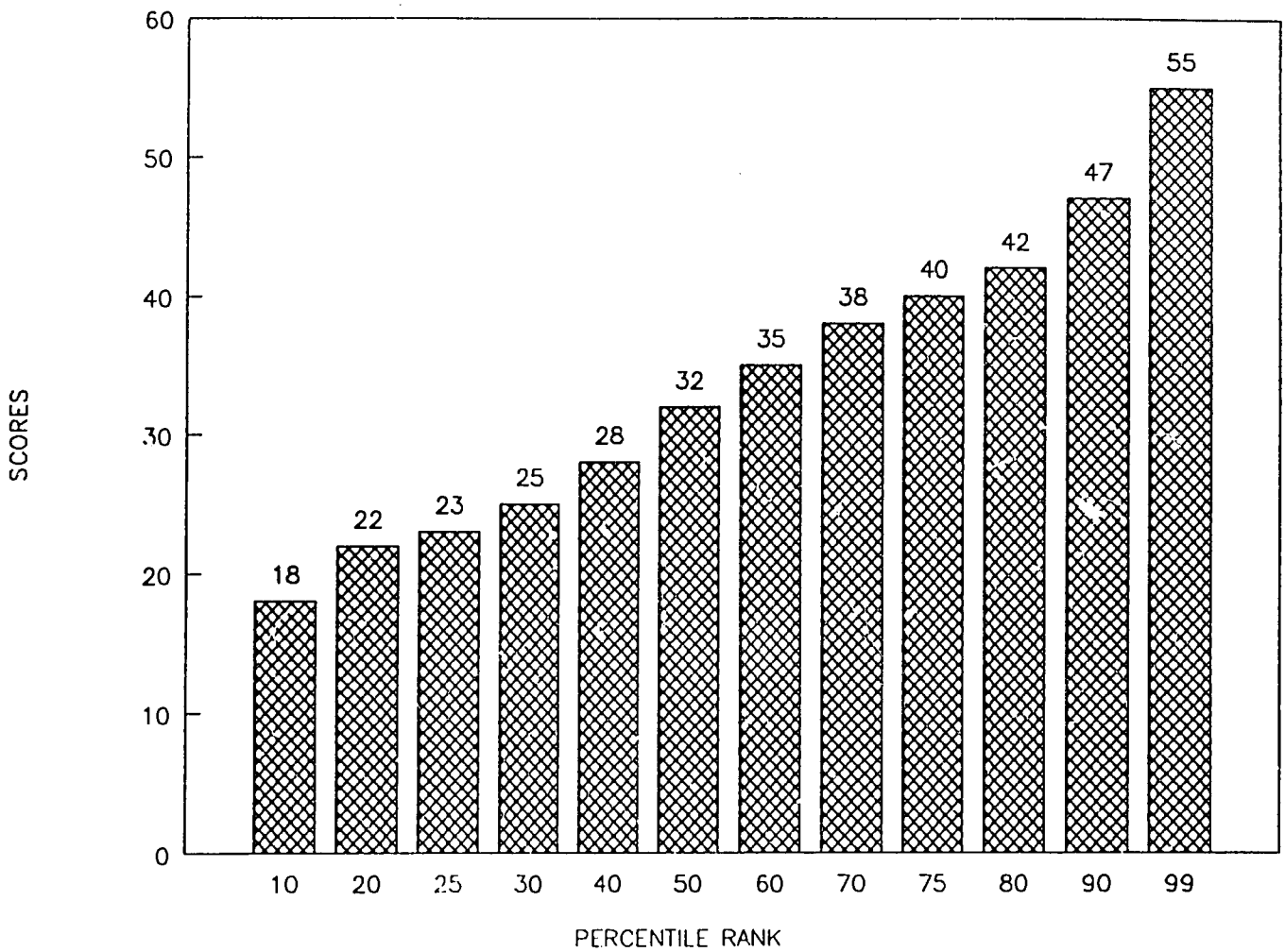
ITEM NO/CONCEPT	*P-VALUE*	*ITEM NO/CONCEPT*	*P-VALUE*
53. Area (Square)	28	57. Bar Graphs (Interpretation)	76
54. Area (Rectangle)	23	58. Bar Graphs (Interpretation)	85
55. Area (Word Problem)	46	59. Line Graphs (Interpretation)	66
56. Volume (Cubic)	30	60. Line graphs (interpretation)	67

OVERALL MEAN (NWFP)=31.84, PERCENT= 53, S.D=11.06, N=1685

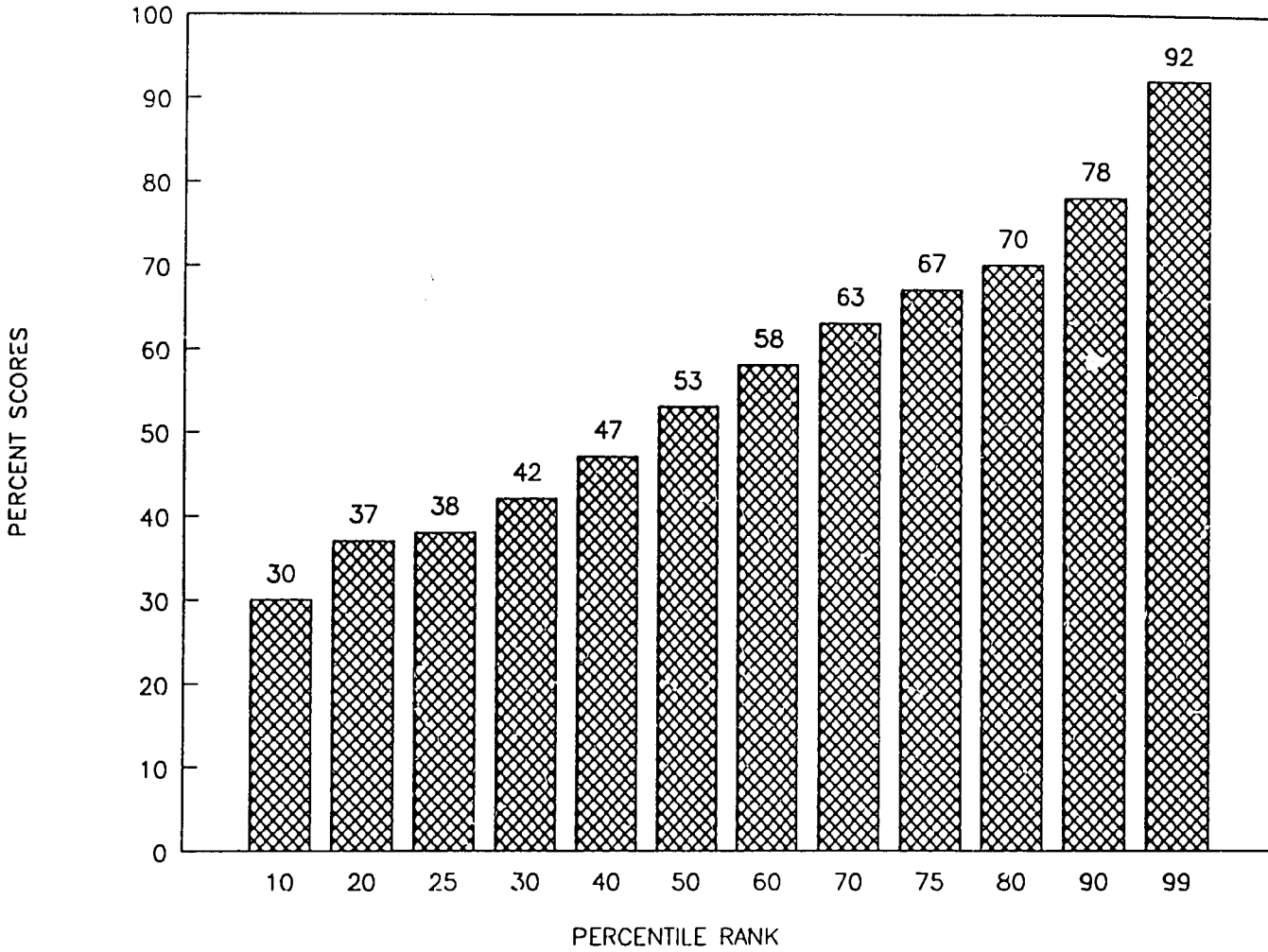
Note : Item no 21 should not be considered due to printing error in the stem of the item.

[K] indicates the total number of items in the test.

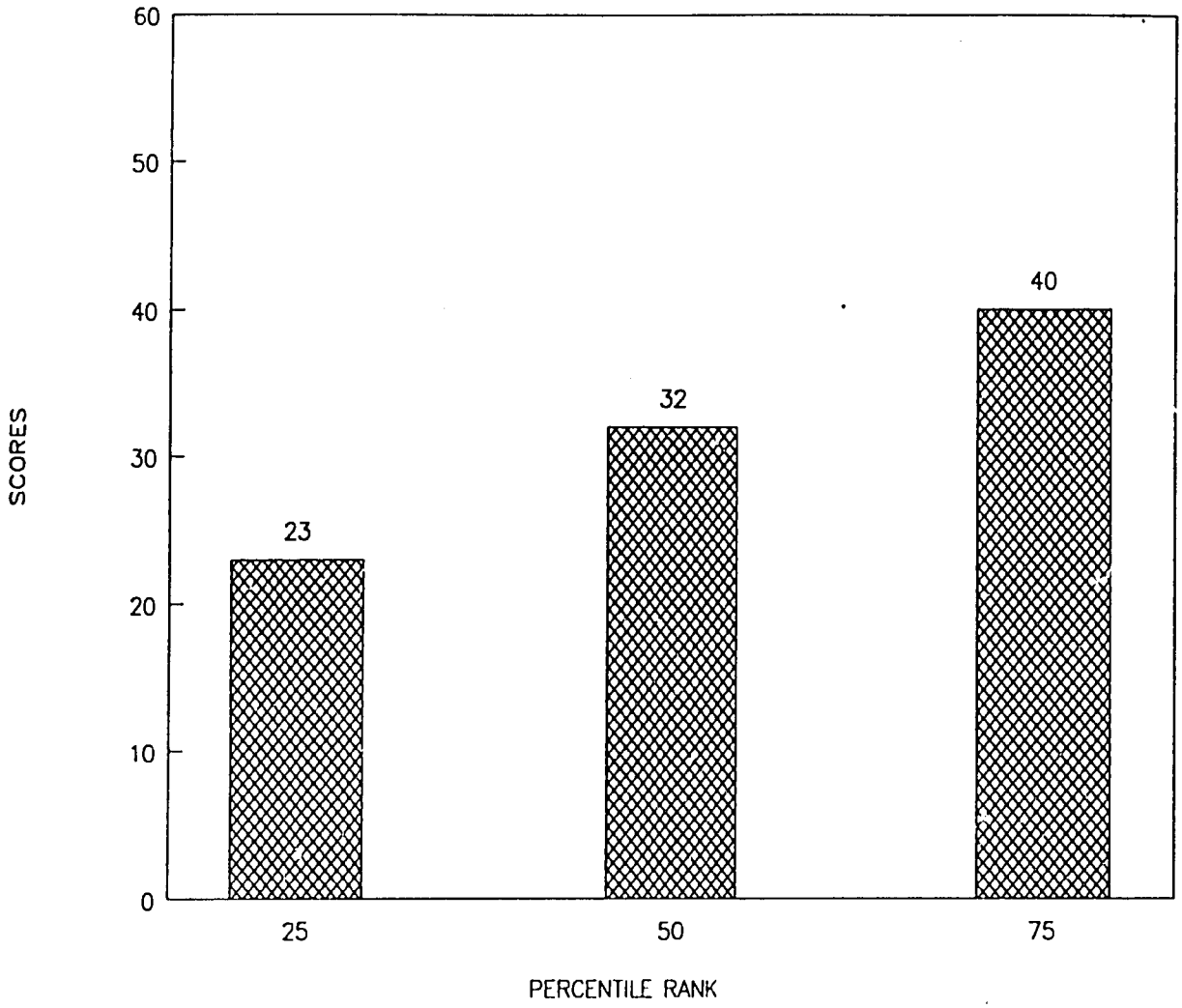
PTC MATH SCORES : N.W.F.P

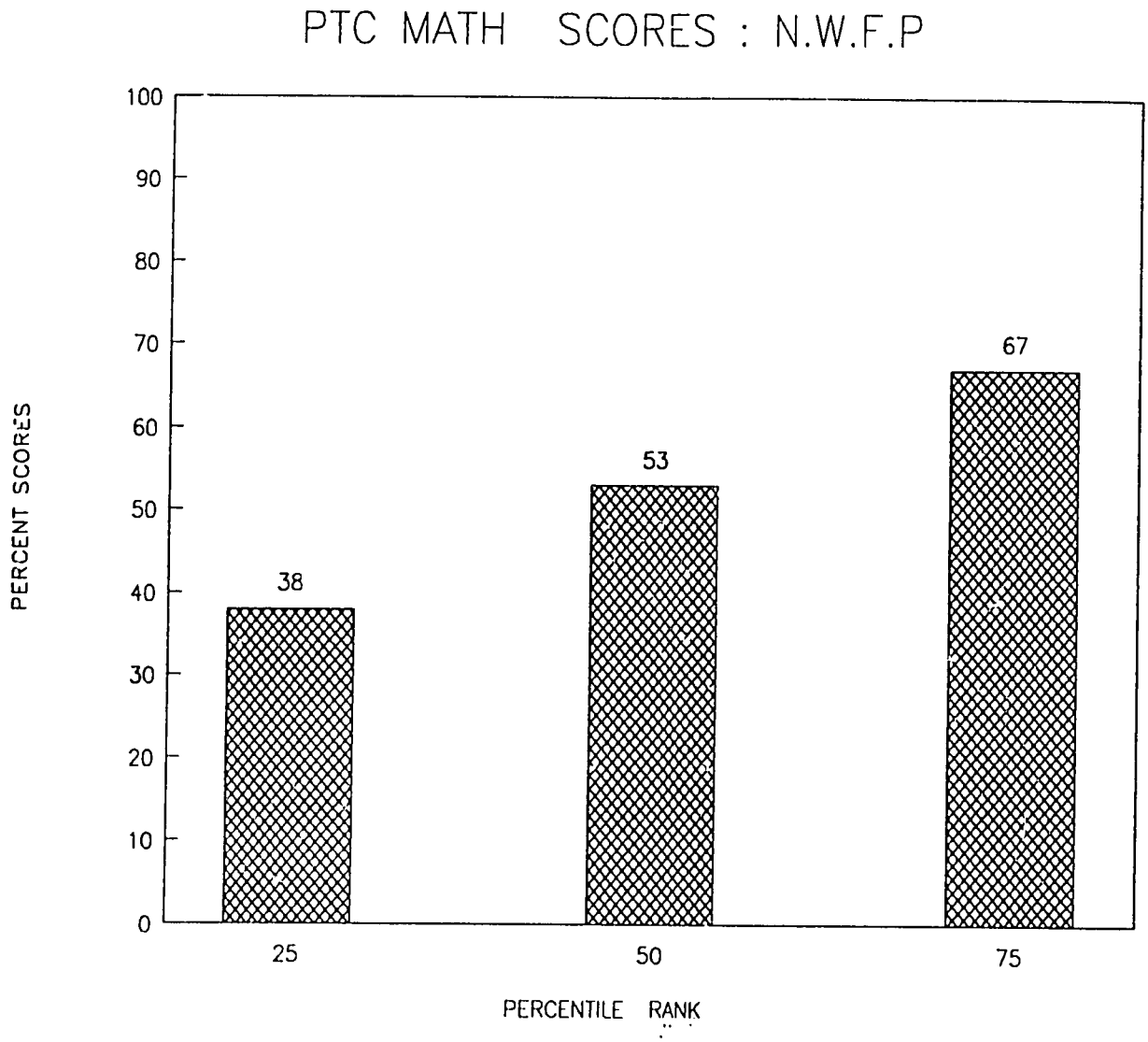


PTC MATH SCORES : N.W.F.P

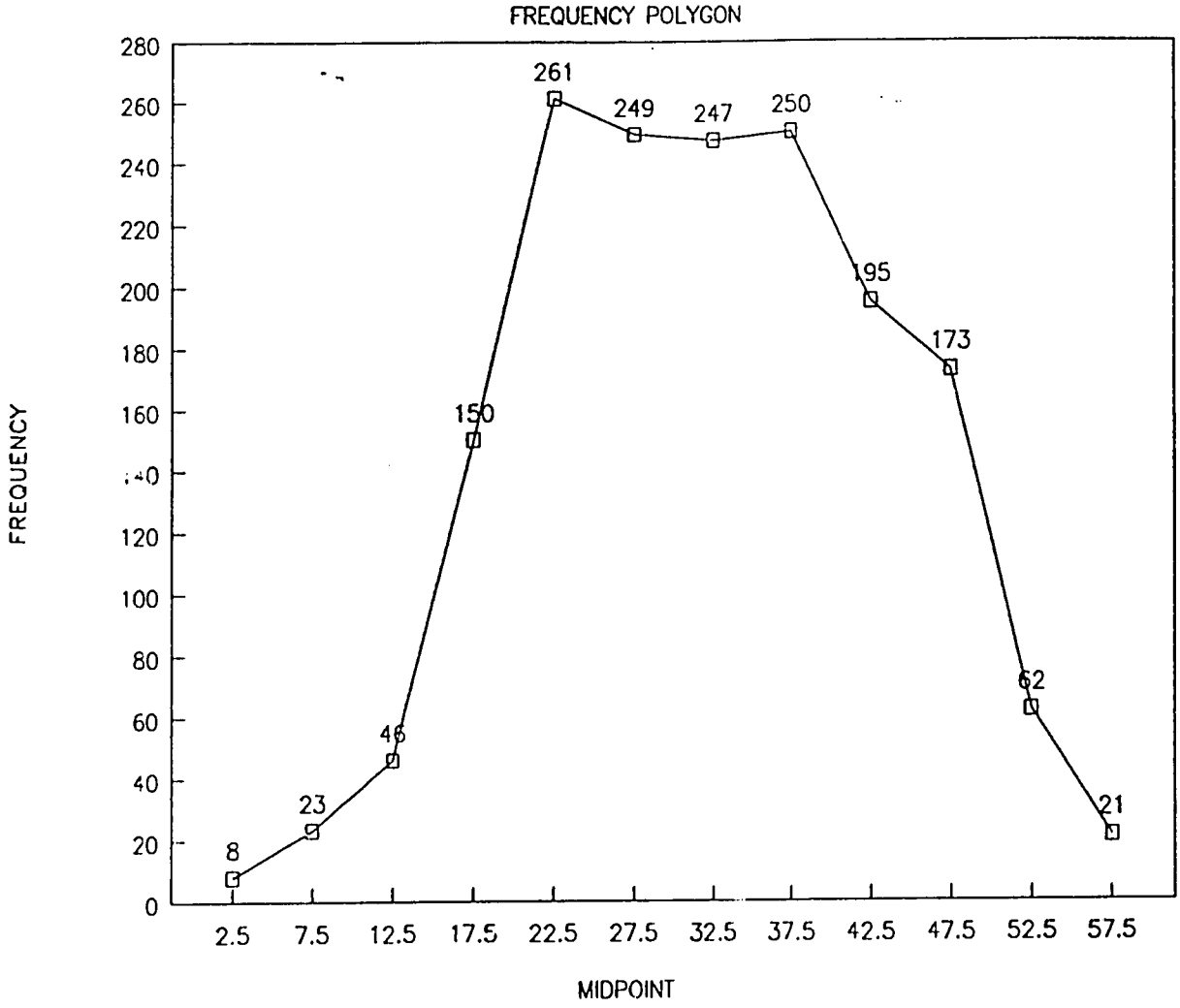


PTC MATH SCORES : N.W.F.P




PTC MATH SCORES : N.W.F.P
PERCENT SCORES
100
90
80
70
60
50
40
30
20
10
0
38
53
67
25
50
75
PERCENTILE RANK

PTC STUDENTS MATHS SCORES : N.W.F.P



Summary of Items-Analysis Results for

Maths for PTC Students

Province N.W.F.P
Pakistan

Subject	**Maths**
No of Items	**60**
No of Students	**1685**
Mean	**31.84**
Variance	**122.31**
Std.Dev	**11.06**
Skew	**-0.001**
Kurtosis	**-0.62**
Minimum	**0.00**
Maximum	**58.00**
Median	**32.00**
Alpha	**0.91**
SEM	**3.27**
Mean P	**0.53**
Meam Item-Tot.	**0.40**
Mean Biserial	**0.53**

TABLE # 1

PERCENT SUCCESS-RATE ON SCIENCE ITEMS (K=60) BY PROVINCE (N.W.F.P) PTC STUDENTS

ITEM NO/CONCEPT	P-VALUE	ITEM NO/CONCEPT	P-VALUE
1. Animals	78	26.Atoms & Molecules	74
2. Animals	78	27.Atoms & Molecules	64
3. Animals	54	28.Atoms & Molecules	45
4. Animals	60	29.Gases	70
5. Animals	64	30.Atoms & Molecules	55
6. Animals	95	31.Atoms & Molecules	60
7. Animals	43	32.Atoms & Molecules	76
8. Animals	39	33.Atoms & Molecules	67
9. Insects	57	34.Water	56
10.Plants	53	35.Atoms & Molecules	48
11.Plants	67	36.Water	78
12.Plants	71	37.Water	35
13.Plants	96	38.Water	29
14.Plants	65	39.Water	28
15.Plants	70	40.Water	14
16.Plants	33	41.Water	70
17.Plants	69	42.Force & Energy	54
18.Plants	39	43.Force & Energy	64
19.Plants	82	44.Force & Energy	34
20.Foods	39	45.Force & Energy	87
21.Foods	63	46.Force & Energy	45
22.Foods	68	47.Force & Energy	38
23.Foods	54	48.Gases	55
24.Foods	37	49.Gases	41
25.Atoms & Molecules	62	50.Light	55

TABLE # 1 (Contd)

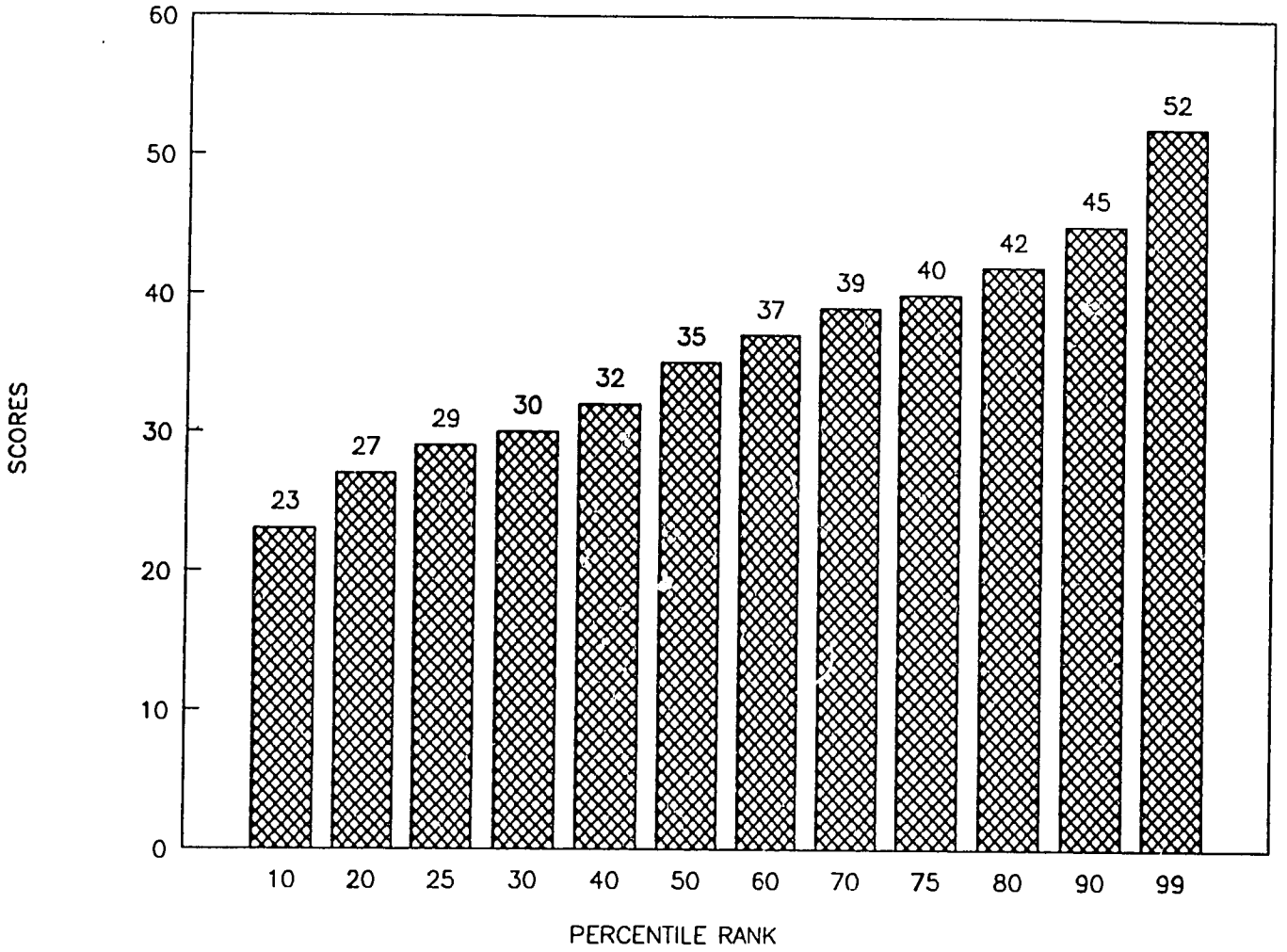
PERCENT SUCCESS-RATE ON SCIENCE ITEMS (K=60)
BY PROVINCE (N.W.F.P)
PTC STUDENTS

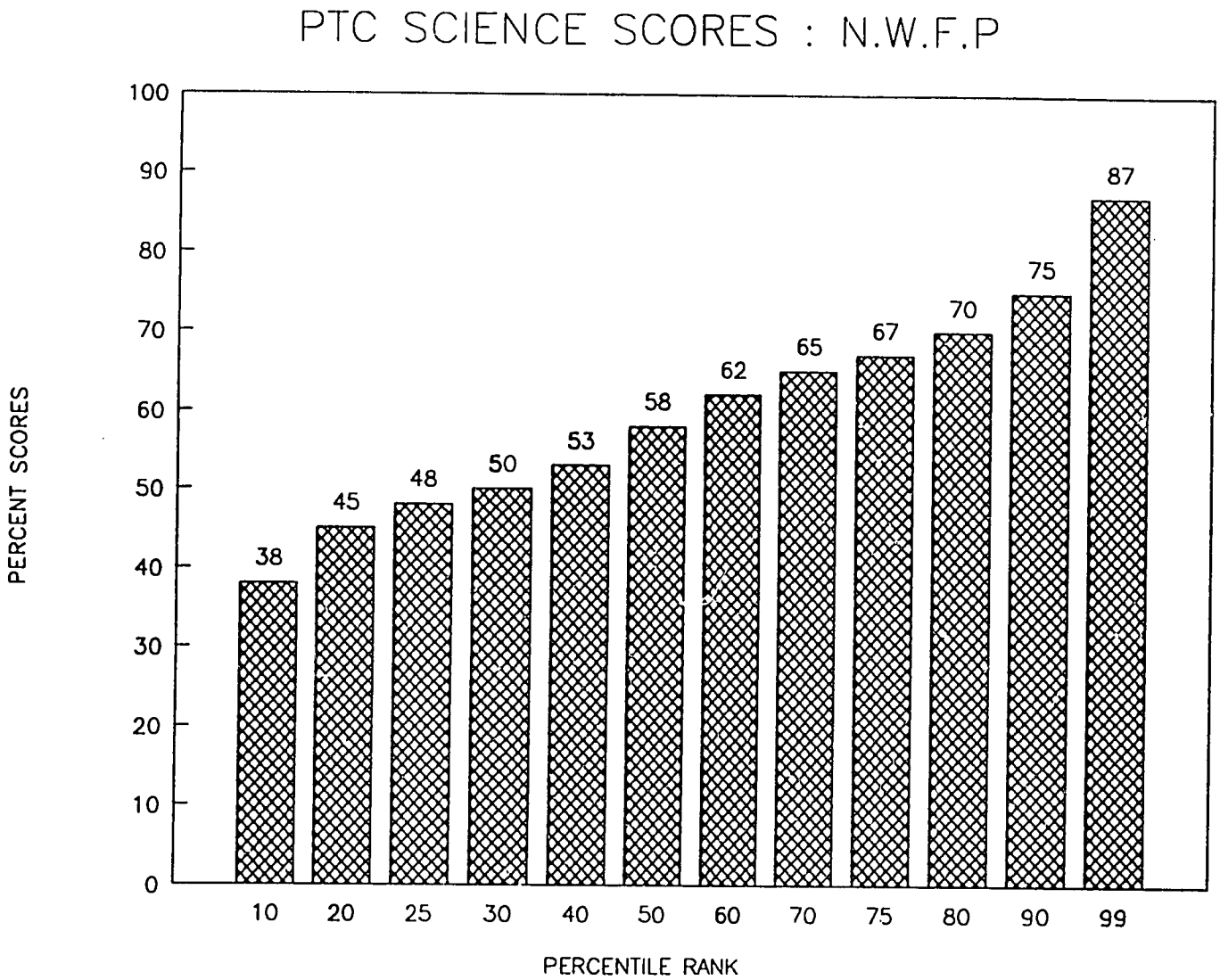
ITEM NO/CONCEPT	P-VALUE	ITEM NO/CONCEPT	P-VALUE
51.Light	46	56.Magnetism	53
52.Light	49	57.Magnetism	78
53.Light	35	58.Magnetism	58
54.Magnetism	70	59.Electro Static	38
55.Magnetism	18	60.Solar System	81

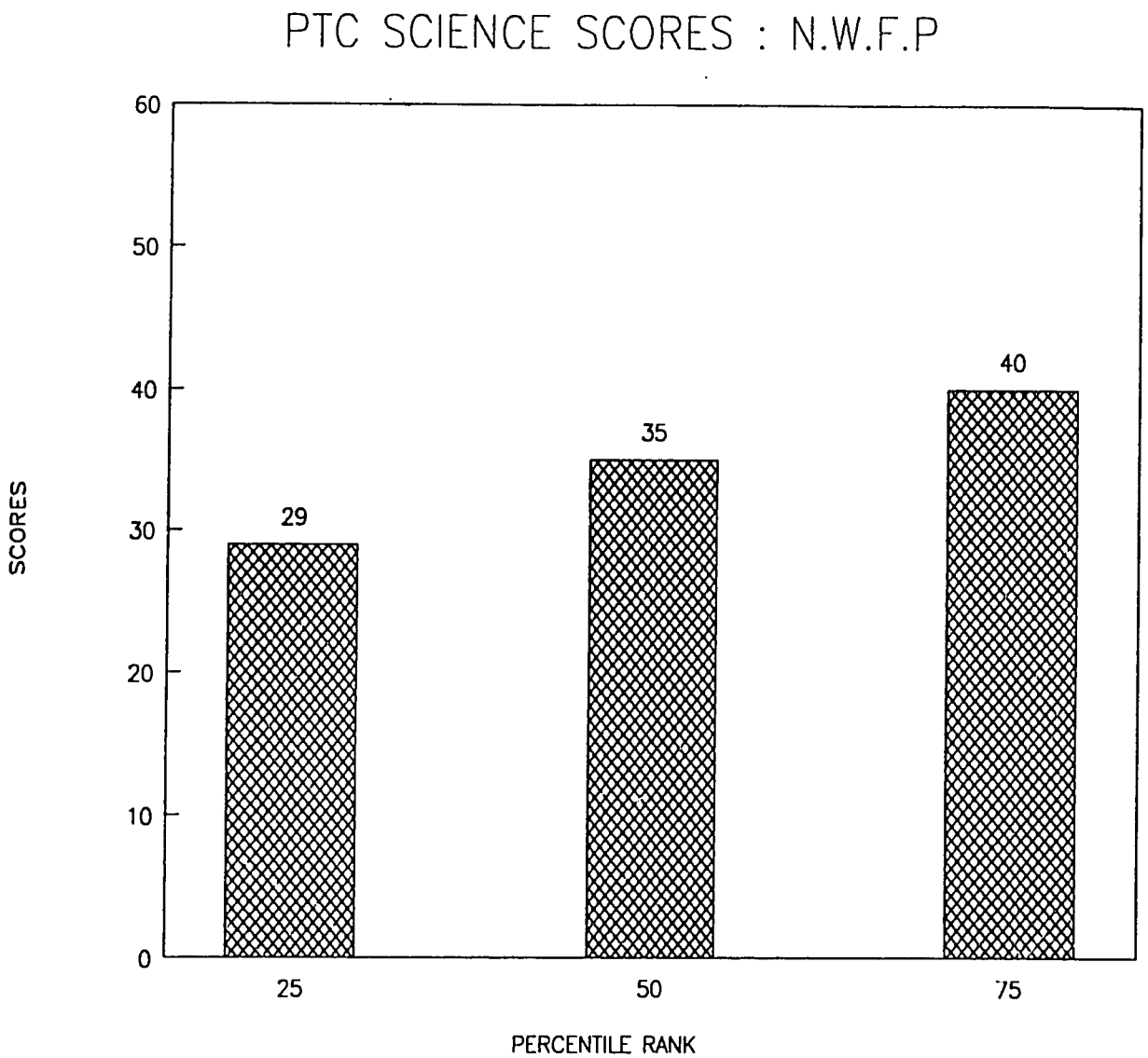
Overall Mean(NWFP)=34.11 ,Percent=57, S.D= 8.78, N= 1684

Note: K indicates the total number of items in the test.

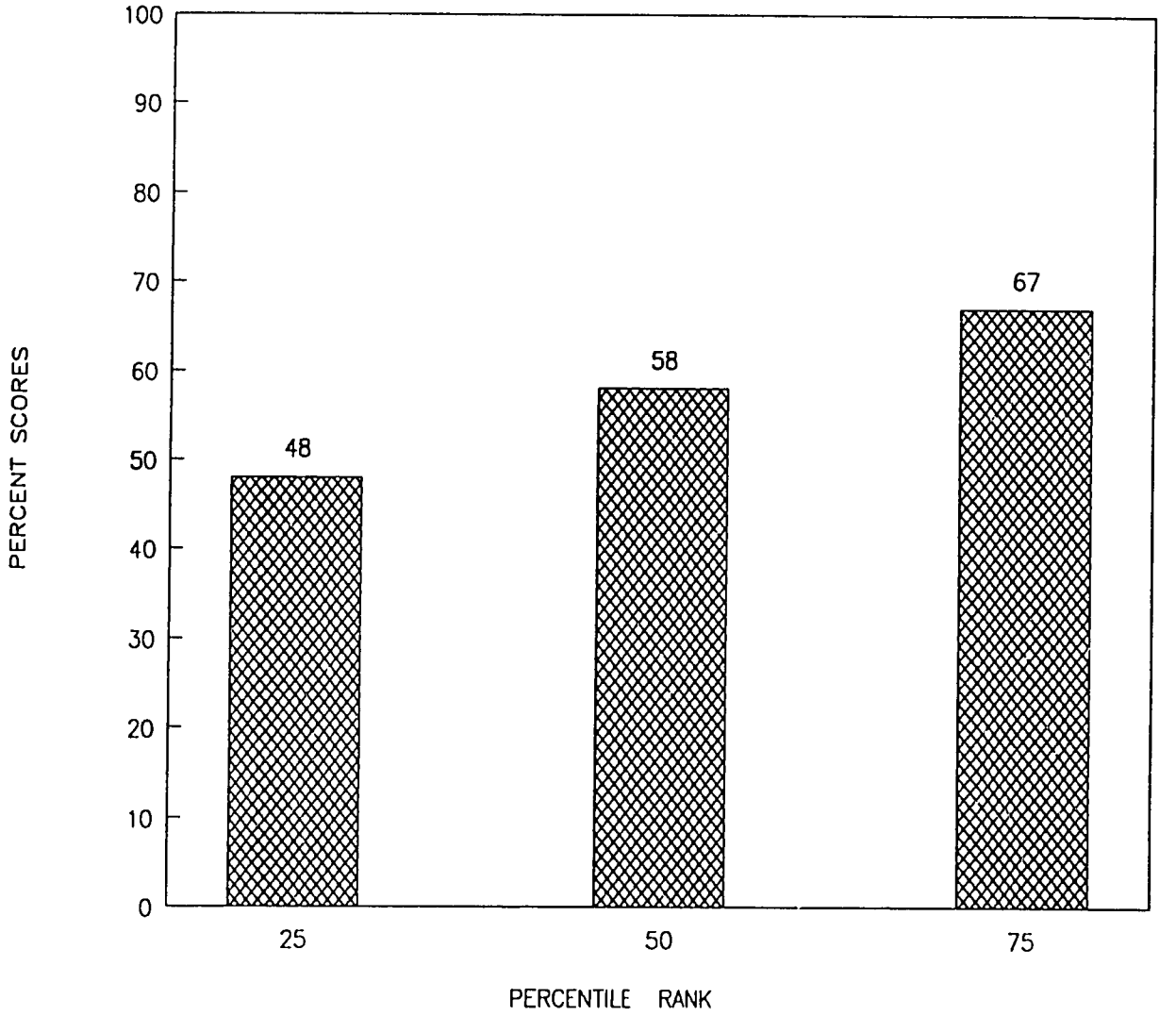
PTC SCIENCE SCORES : N.W.F.P

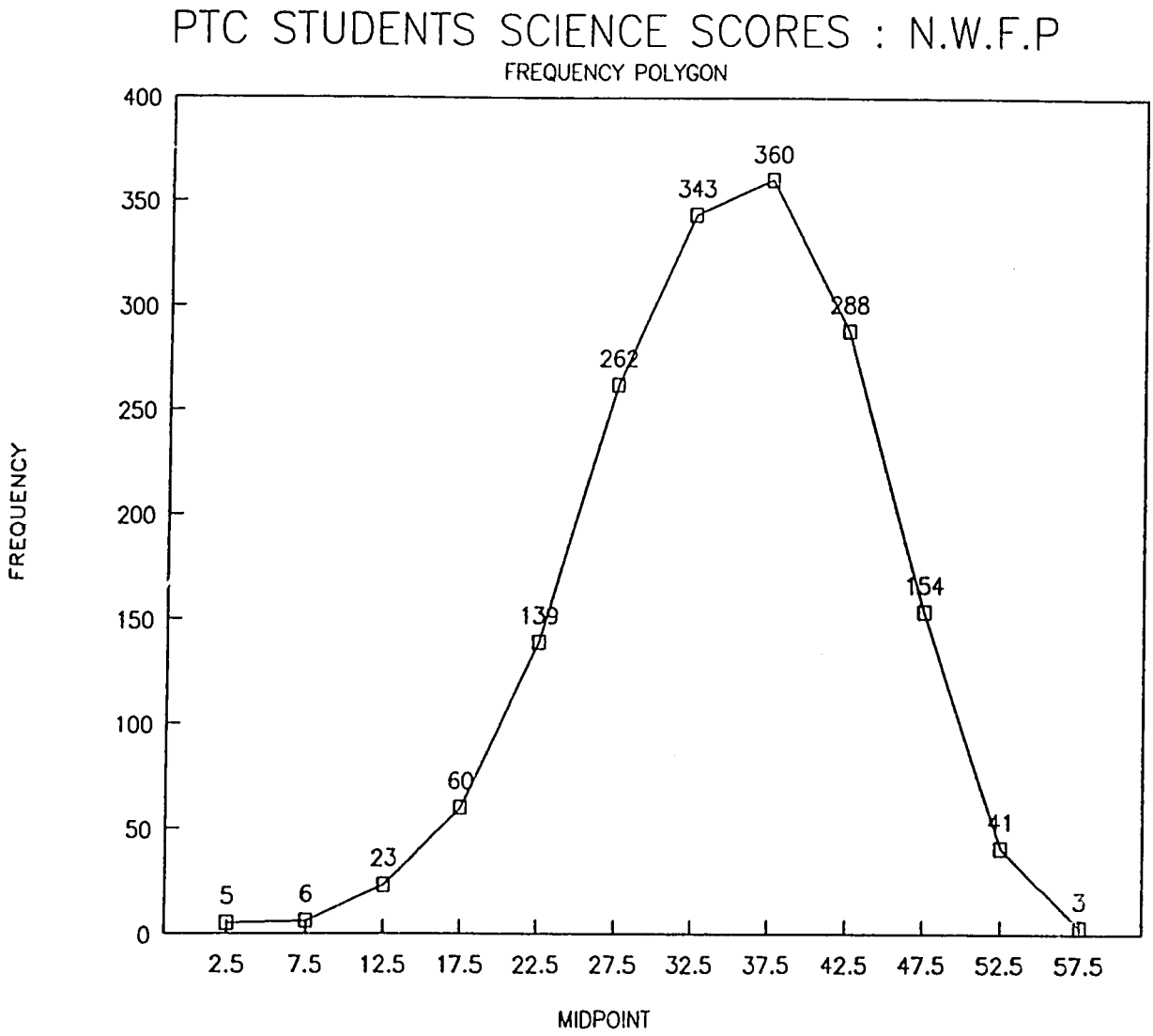



PTC SCIENCE SCORES : N.W.F.P
PERCENT SCORES
100
90
80
70
60
50
40
30
20
10
0
38
45
48
50
53
58
62
65
67
70
75
87
10
20
25
30
40
50
60
70
75
80
90
99
PERCENTILE RANK


PTC SCIENCE SCORES : N.W.F.P
SCORES
60
50
40
30
20
10
0
29
35
40
25
50
75
PERCENTILE RANK

PTC SCIENCE SCORES : N.W.F.P




PTC STUDENTS SCIENCE SCORES : N.W.F.P
FREQUENCY POLYGON
FREQUENCY
400
350
300
250
200
150
100
50
0
5
6
23
60
139
262
343
360
288
154
41
3
2.5
7.5
12.5
17.5
22.5
27.5
32.5
37.5
42.5
47.5
52.5
57.5
MIDPOINT

Summary of Items-Analysis Results for

Science for PTC Students

Province N.W.F.P
Pakistan

Subject	**Science**
No of Items	**60**
No of Students	**1684**
Mean	**34.11**
Variance	**77.04**
Std.Dev	**8.78**
Skew	**-0.35**
Kurtosis	**0.08**
Minimum	**1.00**
Maximum	**56.00**
Median	**35.00**
Alpha	**0.85**
SEM	**3.42**
Mean P	**0.57**
Meam Item-Tot.	**0.32**
Mean Biserial	**0.42**

TABLE # 1

PERCENT SUCCESS-RATE ON URDU ITEMS (K=46) BY PROVINCE (N.W.F.P)

PTC STUDENTS

ITEM NO/CONCEPT	*P-VALUE*	*ITEM NO/CONCEPT*	*P-VALUE*
1.Word Meaning Interpretive Synonyms	77	14.Word Usage Grammar	89
2.Word Meaning Interpretive Synonyms	63	15.Word Usage Grammar	97
3.Word Meaning Interpretive Synonyms	92	16.Word Usage Grammar	97
4.Word Meaning Interpretive Synonyms	53	17.Word Usage Grammar	81
5.Word Meaning Interpretive Synonyms	62	18.Word Usage Grammar	65
6.Word Meaning Interpretive Synonyms	66	19.Word Usage Grammar	90
7.Word Meaning Interpretive Synonyms	47	20.Word Usage Grammar	49
8.Word Meaning Interpretive Antonyms	95	21.Word Usage Grammar	57
9.Word Meaning Interpretive Antonyms	84	22.Word Usage Grammar	88
10.Word Meaning Interpretive Antonyms	80	23.Word Usage Spelling	26
11.Word Meaning Interpretive Antonyms	89	24.Word Usage Spelling	44
12.Word Meaning Interpretive Antonyms	55	25.Word Usage Spelling	75
13.Word Usage Grammar	92	26.Word Usage Spelling	55

TABLE # 1(Contd)

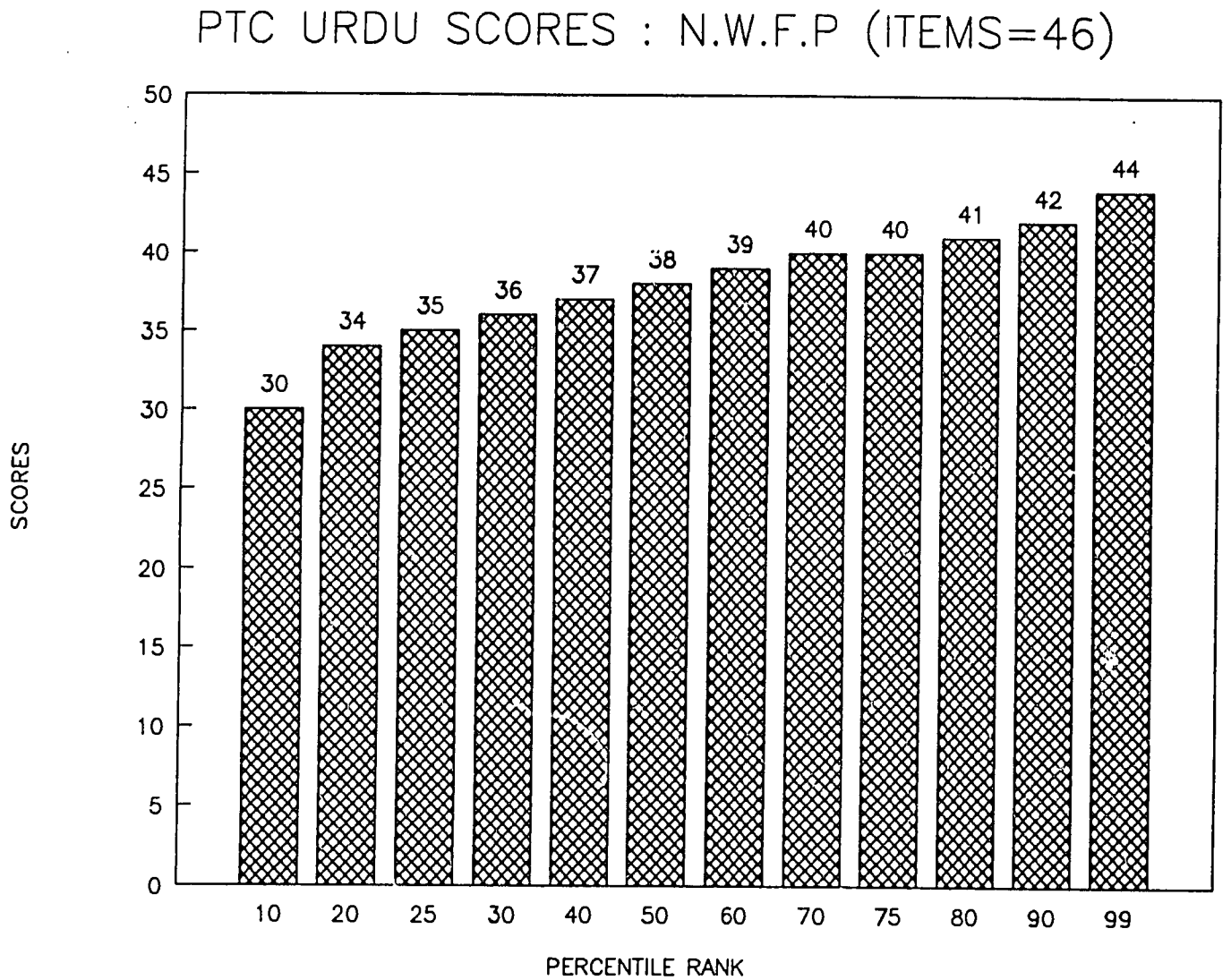
PERCENT SUCCESS-RATE ON URDU ITEMS (K=46) BY PROVINCE (N.W.F.P)

PTC STUDENTS

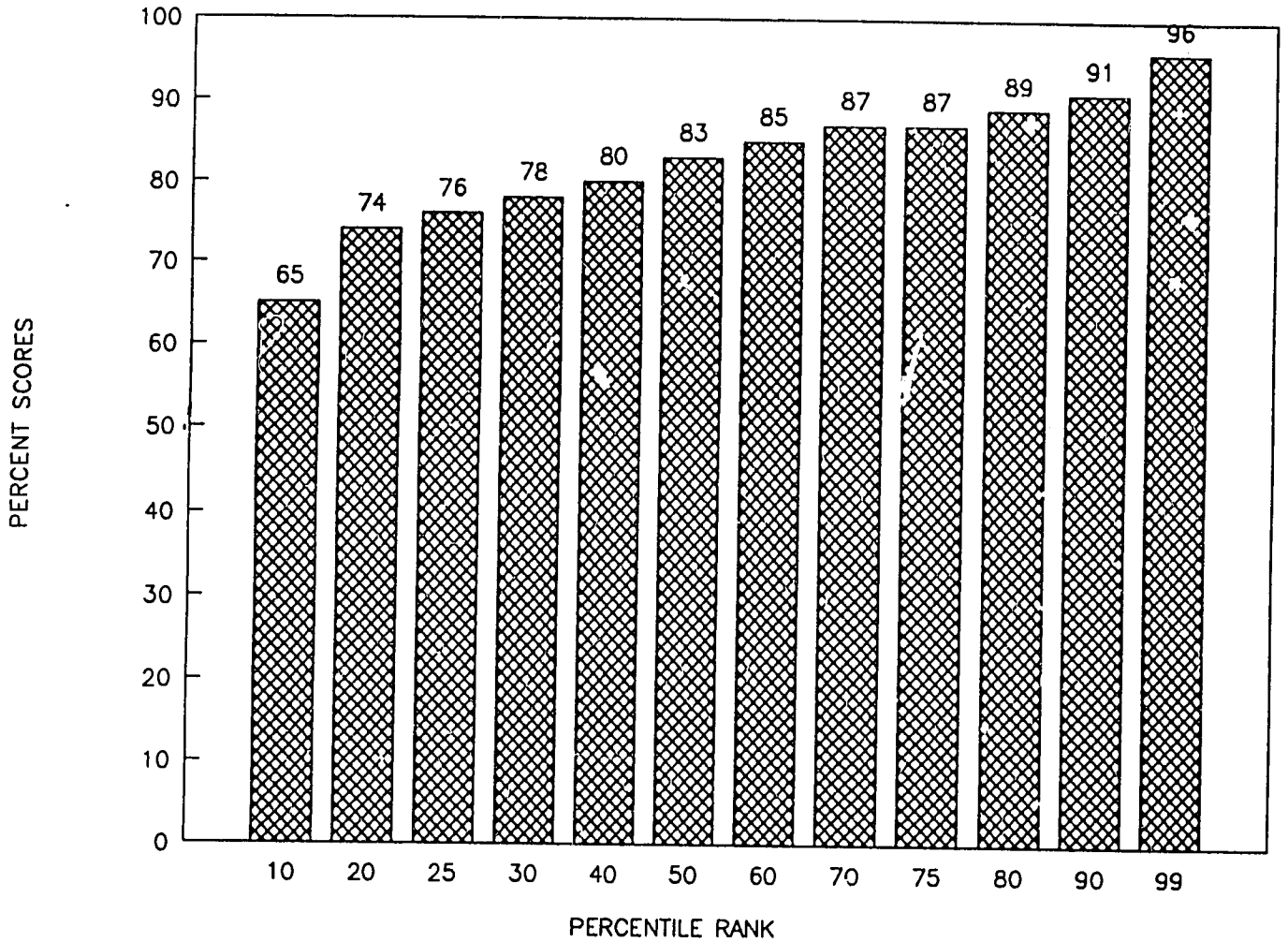
ITEM NO/CONCEPT	*P-VALUE*	*ITEM NO/CONCEPT*	*P-VALUE*
27.Reading Comprehension (Title)	85	37.Sentence Completion (Grammar)	96
28.Deduction	97	38.Sentence Completion (Grammar)	92
29.Deduction	94	39.Sentence Completion (Grammar)	91
30.Reading Comprehension (Title)	53	40.Sentence Completion (Grammar)	86
31.Deduction	94	41.Sentence Completion (Grammar)	97
32.Deduction	88	42.Sentence Completion (Grammar)	96
33.Sentence Completion (Grammar)	94	43.Sentence Completion (Grammar)	92
34.Sentence Completion (Grammar)	96	44.Sentence Completion (Grammar)	87
35.Sentence Completion (Grammar)	93	45.Structure :Meaningful Order	84
36.Sentence Completion (Grammar)	86	46.Structure :Meaningful Order	77

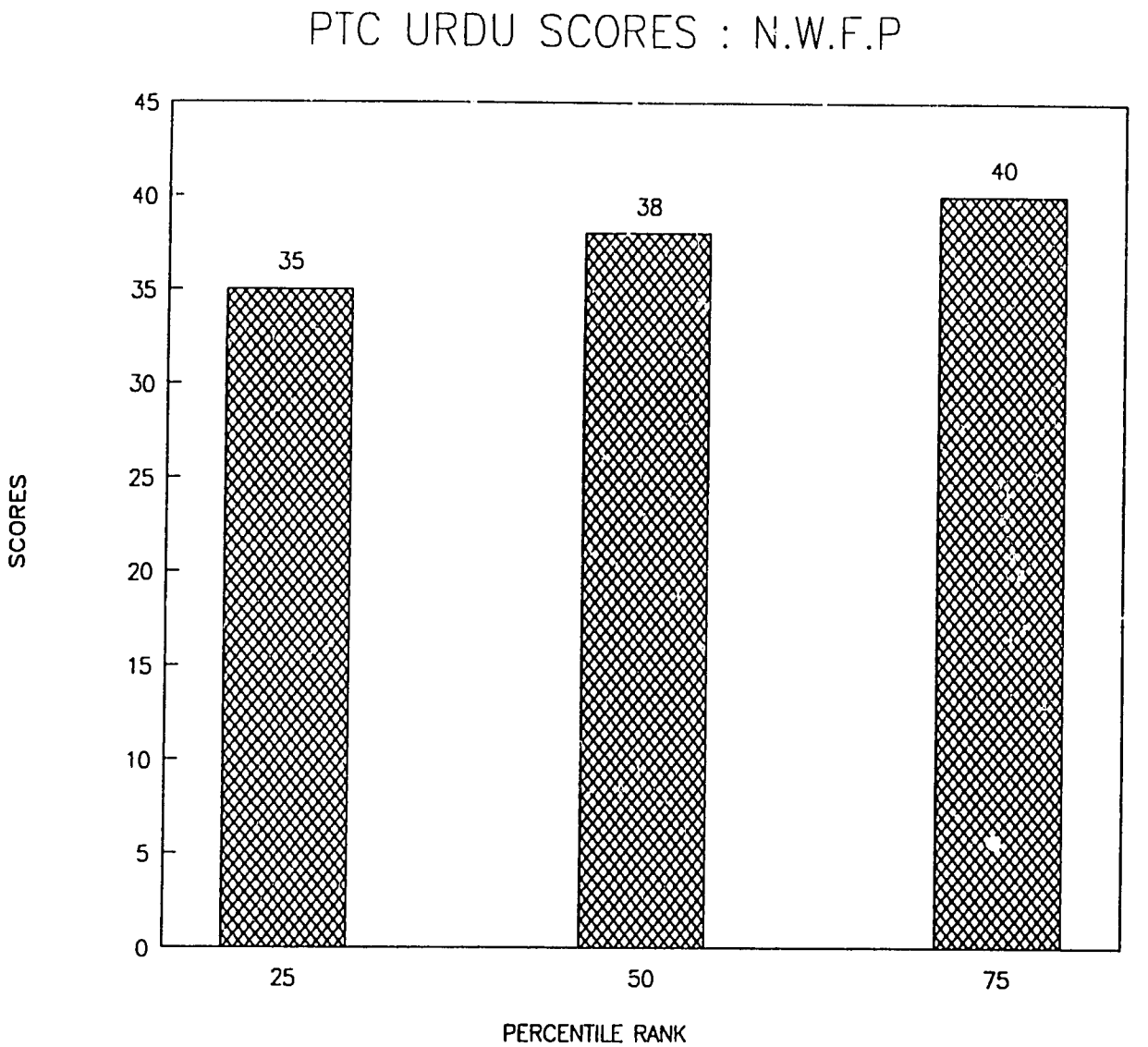
Overall Mean (NWFP)= 36.47, Percent = 79, S.D=5.76, N= 1685

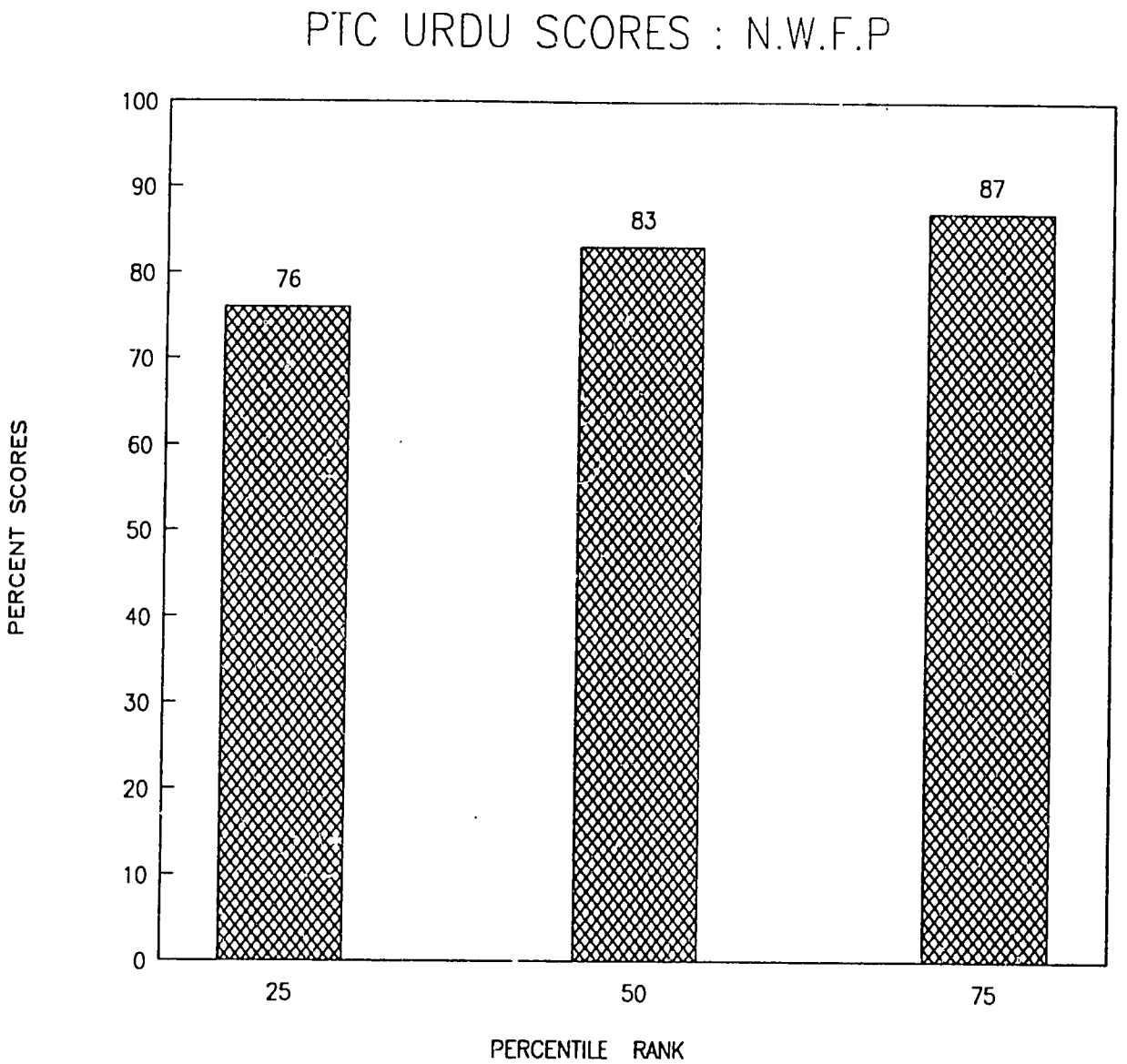
Note: K indicates the total number of items in the test.


PTC URDU SCORES : N.W.F.P (ITEMS=46)
SCORES
50
45
40
35
30
25
20
15
10
5
0
30
34
35
36
37
38
39
40
40
41
42
44
10
20
25
30
40
50
60
70
75
80
90
99
PERCENTILE RANK

PTC URDU SCORES : N.W.F.P

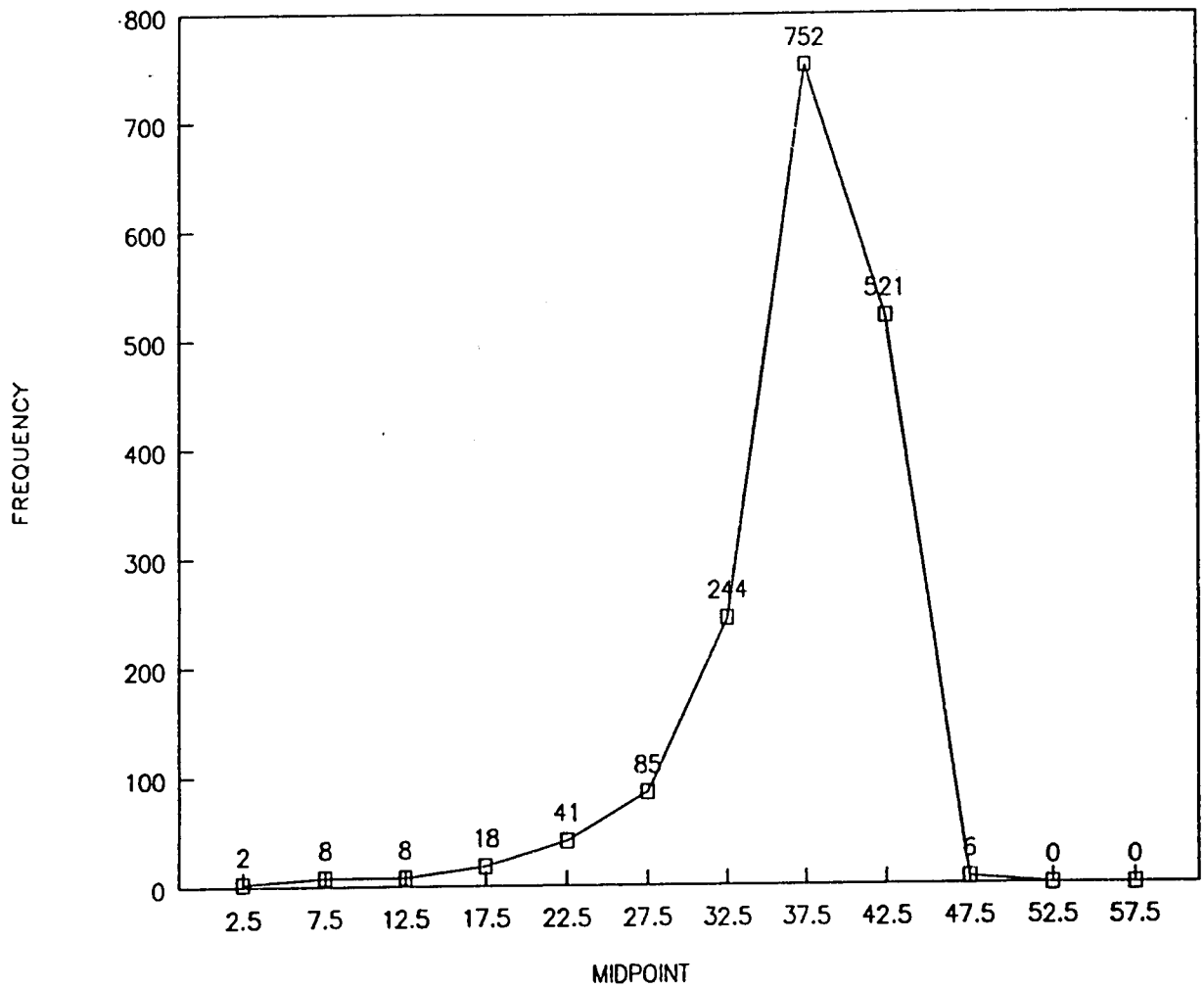



PTC URDU SCORES : N.W.F.P
SCORES
45
40
35
30
25
20
15
10
5
0
35
38
40
25
50
75
PERCENTILE RANK


PTC URDU SCORES : N.W.F.P
100
90
80
70
60
50
40
30
20
10
0
PERCENT SCORES
76
83
87
25
50
75
PERCENTILE RANK

PTC STUDENTS URDU SCORES : N.W.F.P

FREQUENCY POLYGON



Summary of Items-Analysis Results for

Urdu for PTC Students

Province N.W.F.P
Pakistan

Subject	**Urdu**
No of Items	**46**
No of Students	**1685**
Mean	**36.47**
Variance	**33.17**
Std.Dev	**5.76**
Skew	**-2.13**
Kurtosis	**6.47**
Minimum	**0.00**
Maximum	**46.00**
Median	**38.00**
Alpha	**0.83**
SEM	**2.34**
Mean P	**0.79**
Meam Item-Tot.	**0.39**
Mean Biserial	**0.66**

TABLE # 1
PERCENT SUCCESS-RATE ON PASHTO ITEMS (K=50) BY PROVINCE (N.W.F.P)
PTC STUDENTS

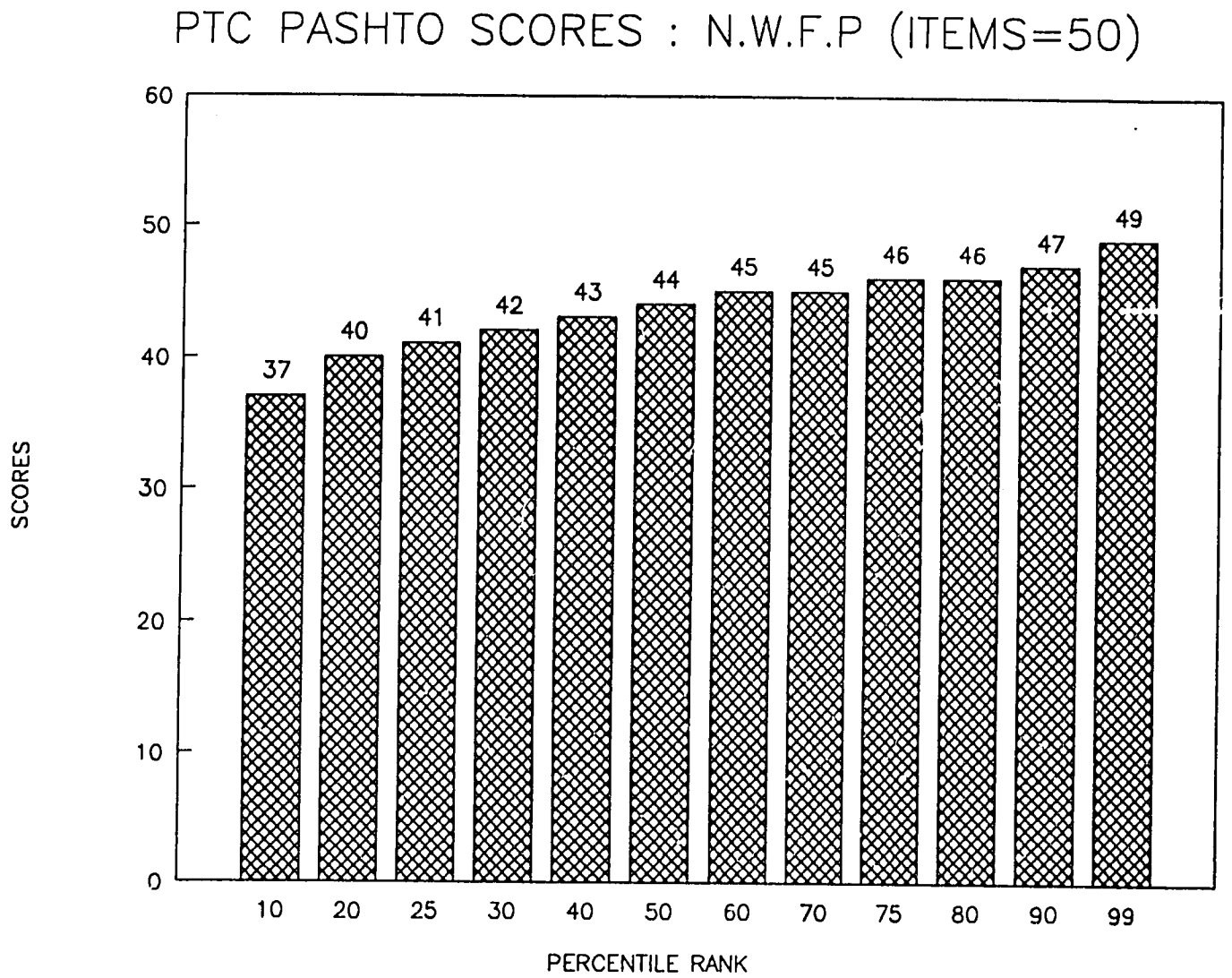
ITEM NO/CONCEPT	*P-VALUE*	*ITEM NO/CONCEPT*	*P-VALUE*
1.Word Meaning Interpretive Synonyms	92	15.Word Usage Grammar	97
2.Word Meaning Interpretive Synonyms	78	16.Word Usage Grammar	97
3.Word Meaning Interpretive Synonyms	89	17.Word Usage Grammar	97
4.Word Meaning Interpretive Synonyms	84	18.Word Usage Grammar	71
5.Word Meaning Interpretive Antonyms	97	19.Word Usage Grammar	81
6.Word Meaning Interpretive Antonyms	98	20.Word Usage Grammar	96
7.Word Meaning Interpretive Antonyms	95	21.Word Usage Grammar	92
8.Word Meaning Interpretive Antonyms	91	22.Word Usage Grammar	95
9.Word Meaning Interpretive Antonyms	80	23.Word Usage Pronoun	93
10.Sentence completion (Reasoning)	89	24.Word Usage Pronoun	95
11.Sentence completion (Reasoning)	96	25.Word Usage Pronoun	96
12.Sentence completion (Reasoning)	96	26.Word Usage Pronoun	92
13.Sentence completion (Reasoning)	84	27.Word Usage Pronoun	93
14.Word Usage Grammar	98	28.Word Usage Pronoun	90

TABLE # 1 (Contd)
PERCENT SUCCESS-RATE ON PASHTO ITEMS (K=50) BY PROVINCE (N.W.F.P)
PTC STUDENTS

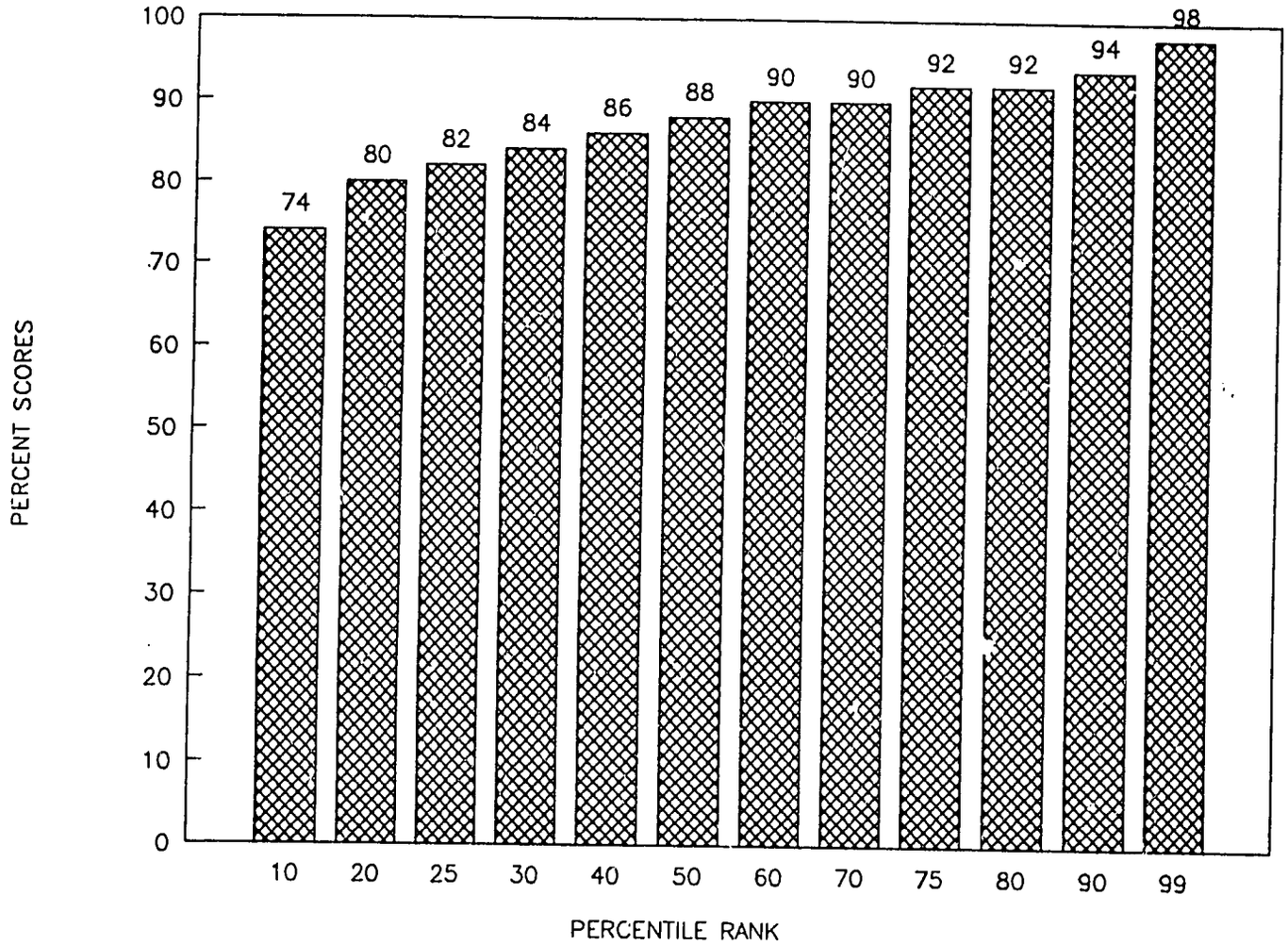
ITEM NO/CONCEPT	*P-VALUE*	*ITEM NO/CONCEPT*	*P-VALUE*
29.Reading Comprehension (Deduction)	95	40.Word Usage Grammar	73
30.Reading Comprehension (Deduction)	94	41.Word Usage Grammar	89
31.Reading Comprehension (Deduction)	94	42.Word Usage Grammar	92
32.Structure (Meaningful Order)	67	43.Word Usage Grammar	93
33.Structure (Meaningful Order)	59	44.Word Usage Grammar	90
34.Structure (Meaningful Order)	40	45.Word Usage Grammar	95
35.Structure (Meaningful Order)	77	46.Word Usage Grammar	93
36.Word Usage Grammar	94	47.Word Usage Spelling	21
37.Word Usage Grammar	95	48.Word Usage Spelling	32
38.Word Usage Grammar	93	49.Word Usage Spelling	64
39.Word Usage Grammar	93	50.Word Usage Spelling	50

Overall Mean (NWFP)=42.49 Percent=85 S.D=5.63 N=754

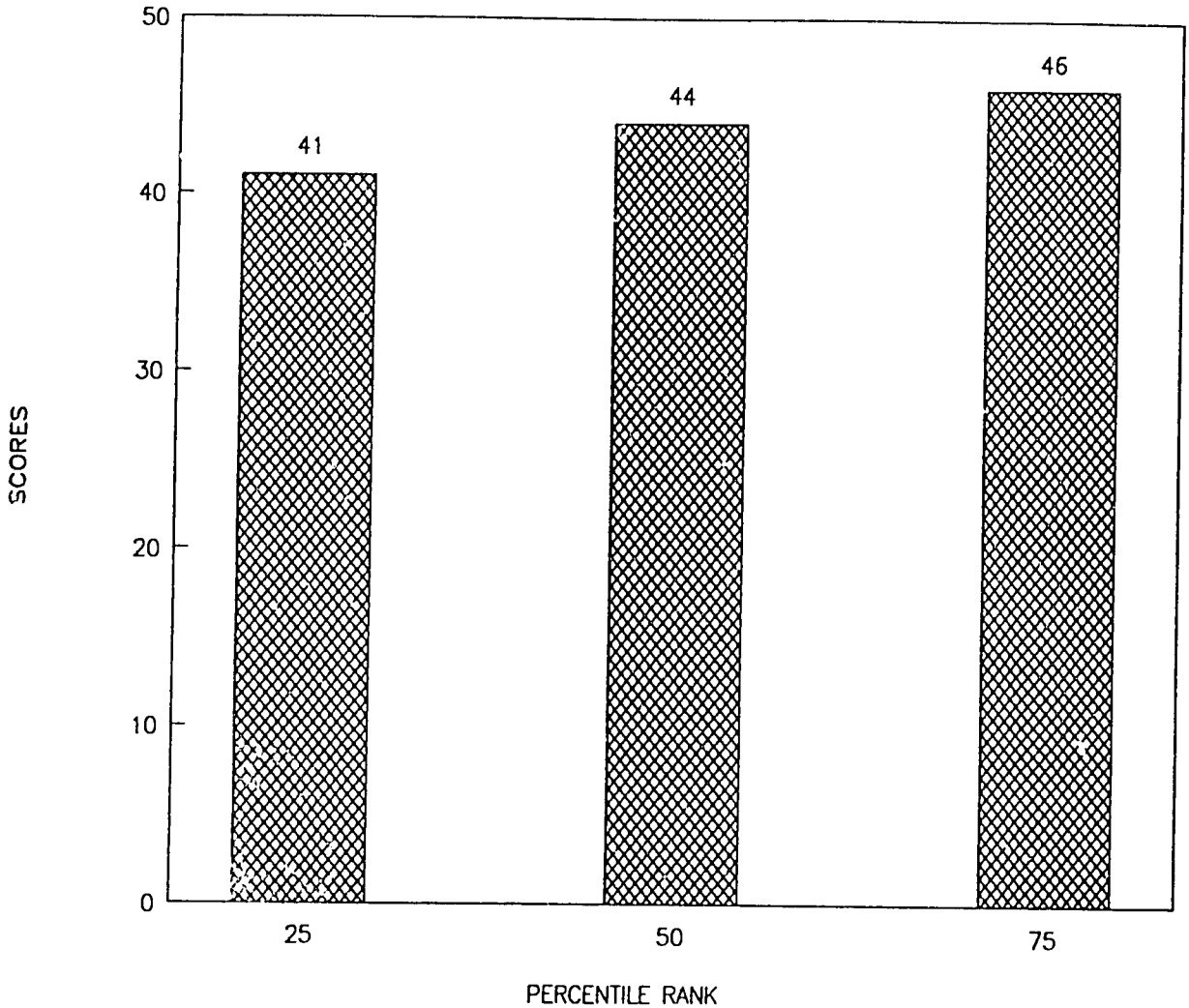
Note: K indicates the total number of items in the test.

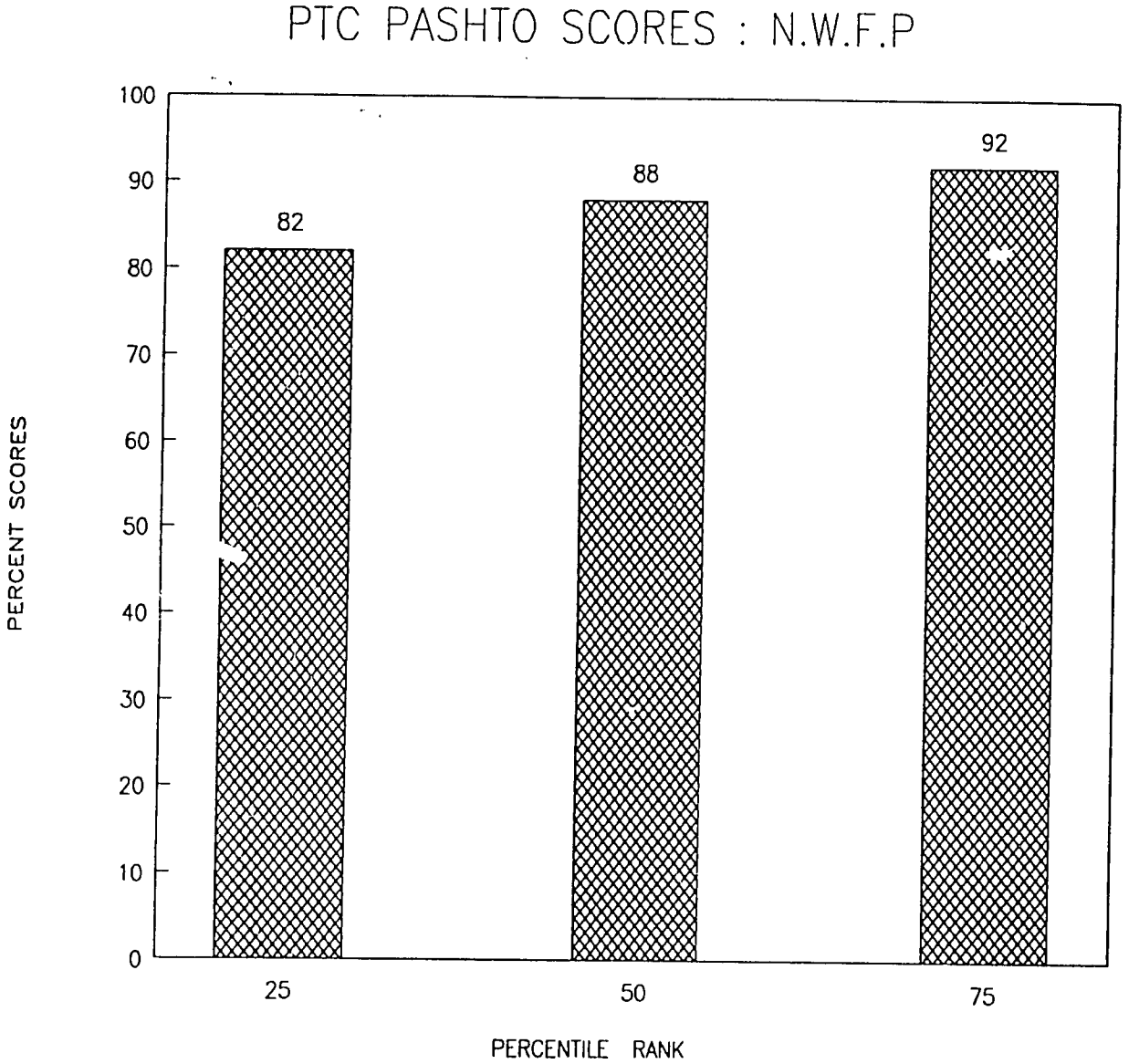

PTC PASHTO SCORES : N.W.F.P (ITEMS=50)
SCORES
60
50
40
30
20
10
0
37
40
41
42
43
44
45
45
46
46
47
49
10
20
25
30
40
50
60
70
75
80
90
99
PERCENTILE RANK

PTC PASHTO SCORES : N.W.F.P



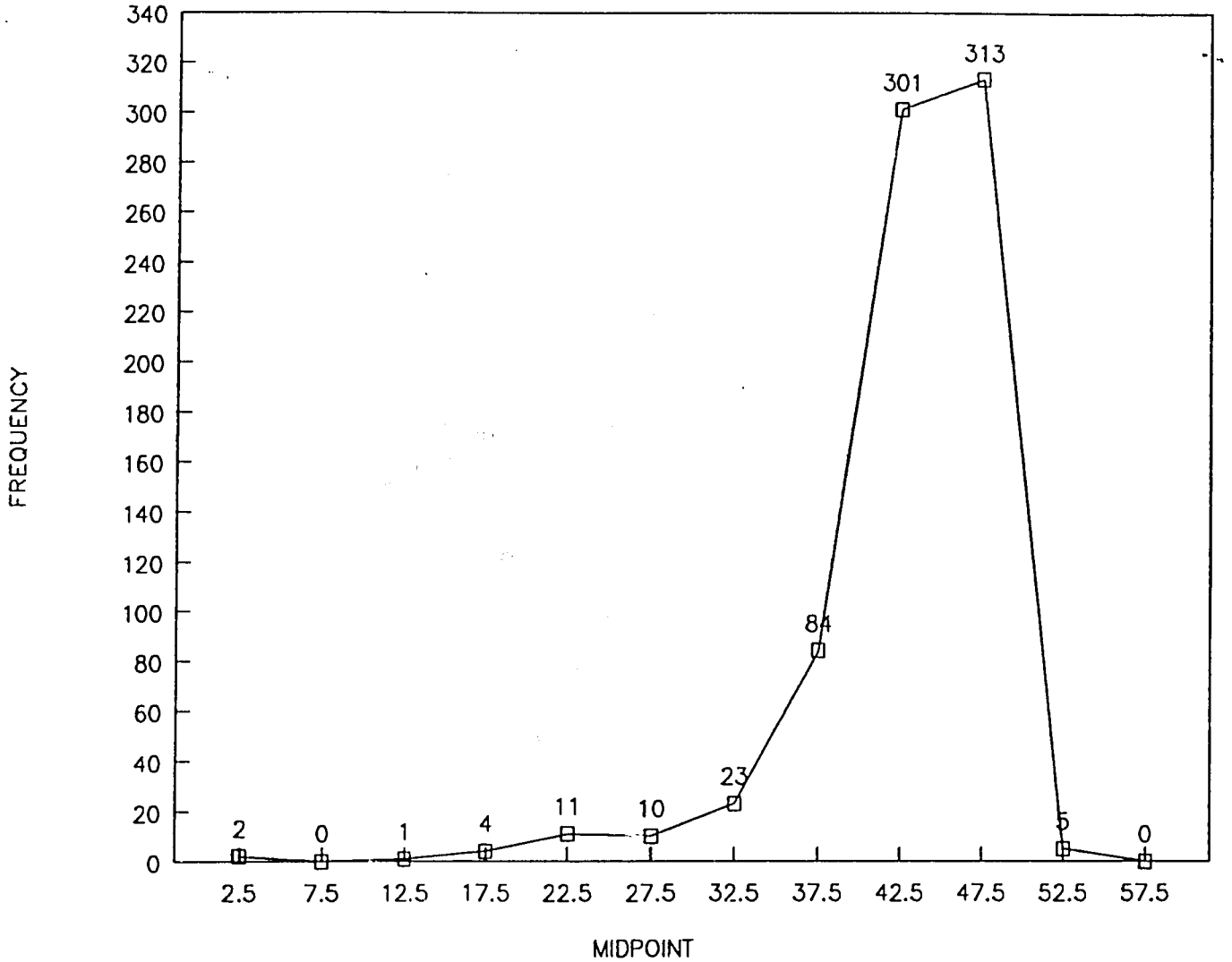
PTC PASHTO SCORES : N.W.F.P




PTC PASHTO SCORES : N.W.F.P
100
90
80
70
60
50
40
30
20
10
0
PERCENT SCORES
82
88
92
25
50
75
PERCENTILE RANK

PUSHTO SCORES OF PTC STUDENTS : N.W.F.P

FREQUENCY POLYGON



Summary of Items-Analysis Results for Pashto for PTC Students Province N.W.F.P (Pakistan)

Subject	*Pashto*
No of Items	*50*
No of Students	*754*
Mean	*42.49*
Std.Dev	*5.63*
Skew	*-2.77*
Kurtosis	*12.33*
Minimum Score	*0.00*
Maximum Score	*50.00*
Median	*44.00*
Alpha	*0.86*
Mean P	*0.85*
Mean Item-Tot.	*0.40*
Mean Biserial	*0.74*

PTC STUDENTS :N.W.F.P

S.NO	COLLEGE NAME	Maths			Scien			Urdu			Pashto		
		$\overline{X}$	R	%	$\overline{X}$	R	%	$\overline{X}$	R	%	$\overline{X}$	R	%
1	Barikot (M)	37.45	1	62	36.70	6	61	36.57	11	79	73.37	3	87
2	Mathra(M)	37.03	2	62	38.33	3	64	38.24	4	83	43.24	5	86
3	Gulbahar In-Ser (M)	37.02	3	62	39.30	2	66	38.77	2	84	44.17	1	88
4	Gulbahar(M)	36.83	4	61	39.60	1	66	38.36	3	83	43.53	2	87
5	Ghoriwala(M)	36.62	5	61	37.06	5	62	36.66	10	80	-----	---	---
6	Karak(M)	36.49	6	61	35.72	8	60	36.18	12	79	-----	---	---
7	Haripur(M)	36.28	7	60	38.17	4	64	39.28	1	85	-----	---	---
8	D.I.Khan(M)	35.83	8	58	36.04	7	60	36.99	9	80	-----	---	---
9	Kohat(M)	32.47	9	54	35.42	9	59	37.79	7	82	-----	---	---
10	Thana(M)	30.61	10	51	32.89	10	55	32.74	16	71	43.30	4	87
11	Charsadda(F)	30.39	11	51	33.29	11	55	38.30	5	83	42.59	6	85
12	Dabgari(F)	30.13	12	50	29.66	17	49	36.10	13	78	42.10	7	84
13	Mansehra(F)	29.73	13	50	31.92	12	53	38.17	6	83	-----	---	---
14	D.I.Khan(F)	28.83	14	48	31.49	13	52	37.53	8	82	-----	---	---
15	Kohat(F)	26.34	15	44	30.53	15	51	35.94	15	78	-----	---	---
16	Dargai(F)	25.32	16	42	31.39	14	52	36.08	14	78	41.97	8	84
17	Drosh(M)	22.58	17	38	29.77	16	50	32.58	17	71	-----	---	---
18	Kh.Khela(F)	19.42	18	32	23.37	18	39	29.10	18	63	38.07	9	7

APPENDIX J:
PERCENT SUCCESS-RATE ON MATH ITEMS FOR BALOCHISTAN

TABLE # 1

PERCENT SUCCESS-RATE ON MATH ITEMS (K=60) BY PROVINCE BALOCHISTAN

ITEM NO/CONCEPT	P-VALUE/Bis		ITEM NO/CONCEPT	P-VALUE/Bis	
1.Equivalence:Roman Arabic Numerals	86	0.24	13.Greatest Common Divisor	68	0.37
2.Prime Number	14	0.28	14.Common Multiples	86	0.58
3.Divisibility	49	0.34	15.Common Multiples	55	0.40
4.Prime Factorization	44	0.51	16.Least Common Multiples (LCM)	45	0.33
5.Place Value: Whole Number	49	0.47	17.LCM (Word Problem)	83	0.31
6.Fractions (Num/Denominator)	83	0.49	18.Associated Property	75	0.56
7.Fractions (Interpration)	35	0.37	19.Fractions (Simplification)	42	0.34
8.Proper/Improper Fractions	49	0.37	20.Mixed Fractions (Simplification)	72	0.60
9.Composite Numbers	60	0.39	21.Fractions (Addition)	78	0.52
10.Common Divisors	47	0.49	22.Fractions (Addition)	22	0.22
11.Equivalent Fractions	36	0.41	23.Fractions (Addition)	12	0.12
12.Fractions (Interpretation)	53	0.30	24.Fractions (Mixed + whole No)	62	0.62

TABLE # 1 (Contd)

PERCENT SUCCESS-RATE ON MATH ITEMS (K=60) BY PROVINCE BALOCHISTAN

ITEM NO/CONCEPT	P-VALUE/Bis		ITEM NO/CONCEPT	P-VALUE/Bis	
25.Fraction (Subtraction)	84	0.84	37.Conversion to Decimal Fractions	25	0.57
26.Fraction (Subtraction)	14	0.14	38.Conersion to Decimal Fractions	18	0.42
27.Mixed Fractions (Subtraction)	28	0.28	39.Conversion to Mixed Fractions	66	0.34
28.Fractions (Multiplication)	72	0.72	40.Decimal Fractions (Place Value)	18	0.12
29.Mixed Fractions (Multiplication)	37	0.51	41.Decimal Fractions (Subtraction)	31	0.63
30.Fractions (Addition)	16	0.47	42.Decimal Fractions(Multiplication)	31	0.48
31.Mixed Fractions (Division)	30	0.58	43.Decimal Fractions(Multiplication)	69	0.44
32.Fractions (word problem)	19	0.21	44.Decimal Fractions (Multiplication)	25	0.46
33.Fractions (Mixed Operations)	33	0.57	45.Decimal Fractions (Division)	16	0.03
34.Fractions (Mixed Operation)	23	0.54	46.Decimal Fractions (Division)	47	0.39
35.Fractions (Order)	33	0.29	47.Decimal Fractions (Division)	15	0.02
36.Decimal Fractions (Addition)	43	0.29	48.Decimal Fractions (Mixed Operation)	29	0.18

TABLE # 1 (Contd)

PERCENT SUCCESS-RATE ON MATH ITEMS (K=60) BY PROVINCE BALOCHISTAN

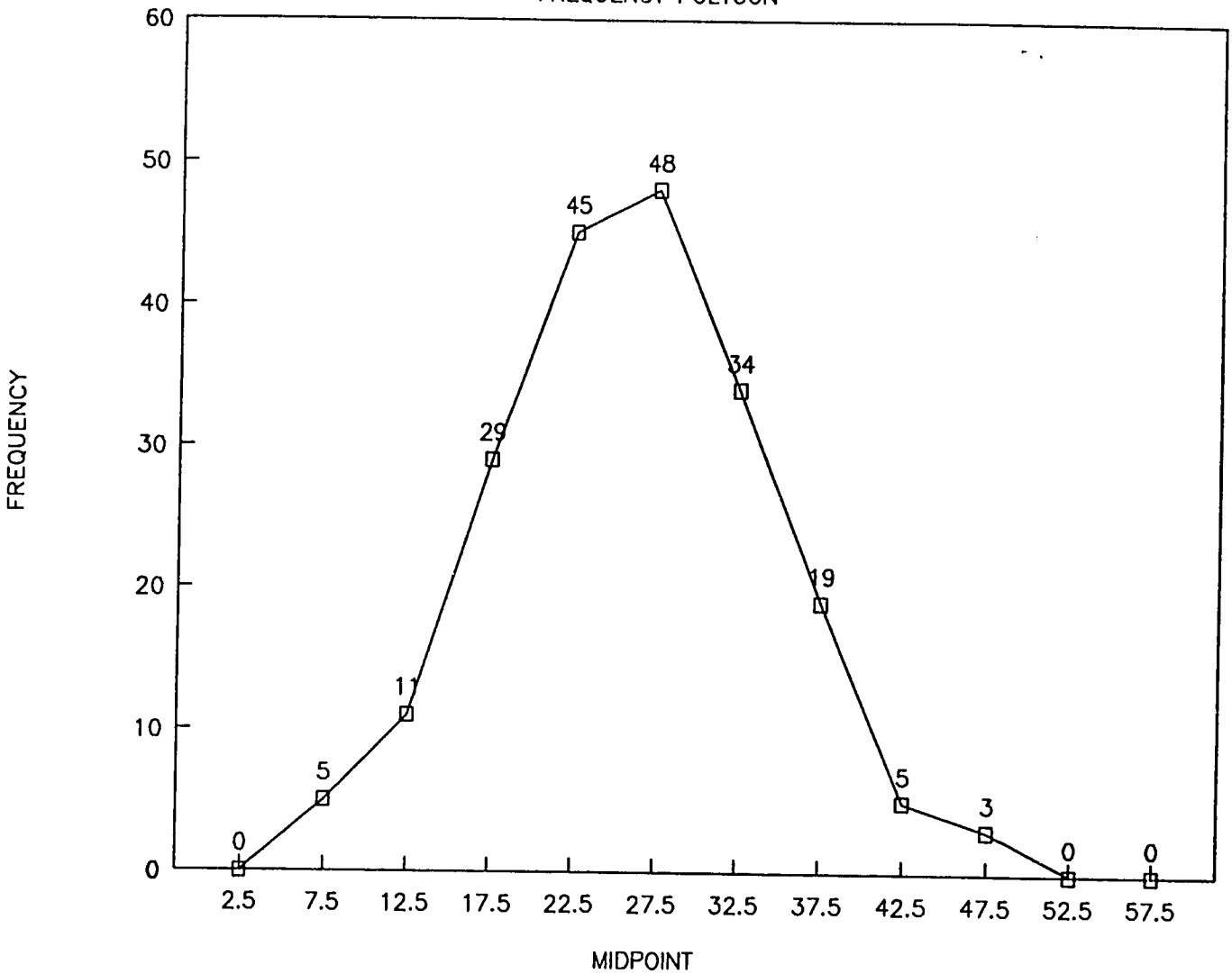
ITEM NO/CONCEPT	P-VALUE/Bis		ITEM NO/CONCEPT	P-VALUE/Bis	
49.Unitary Method	70	0.54	55.Area (Word Problem)	37	0.45
50.Average	40	0.53	56.Volume (Cubic)	13	0.29
51.Rays and Angles	12	0.07	57.Bar Graphs (Interpretation)	70	0.47
52.Area (Triangle)	6	0.21	58.Bar Graphs (Interpretation)	69	0.42
53.Area (Square)	24	0.19	59.Lin Graphs (Interpretation)	39	0.59
54.Area (Rectangle)	25	0.18	60.Lin Graphs (Interpretation)	42	0.53

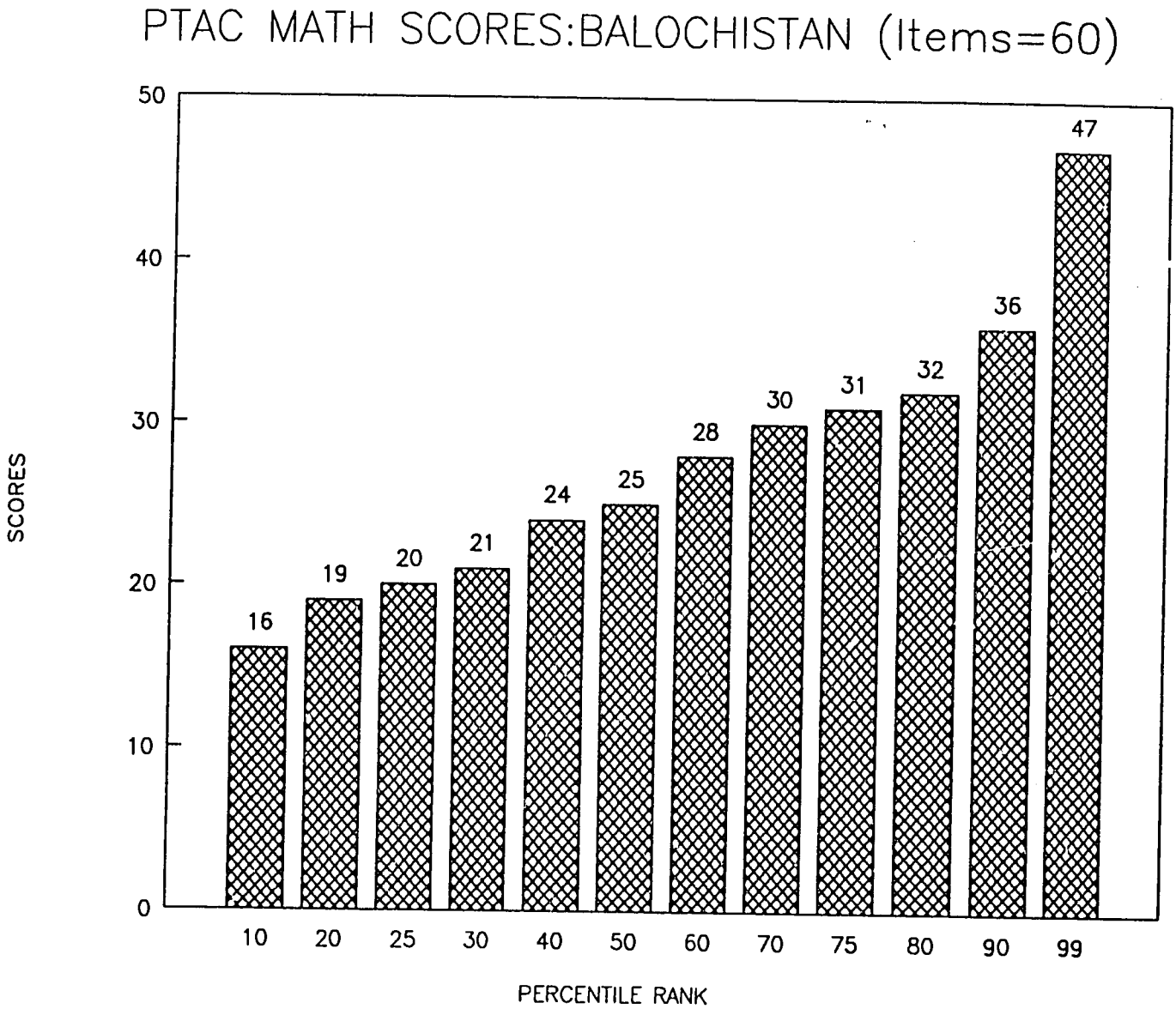
Mean = 25.70, Percent= 43, S.D= 8.09, n= 199

Note: K indicates the total number of items in the test.

PTAC STUDENTS' MATH SCORES: BALOCHISTAN

FREQUENCY POLYGON




PTAC MATH SCORES:BALOCHISTAN (Items=60)
SCORES
50
40
30
20
10
0
16
19
20
21
24
25
28
30
31
32
36
47
10
20
25
30
40
50
60
70
75
80
90
99
PERCENTILE RANK

PTAC MATH SCORES:BALOCHISTAN (Items=60)

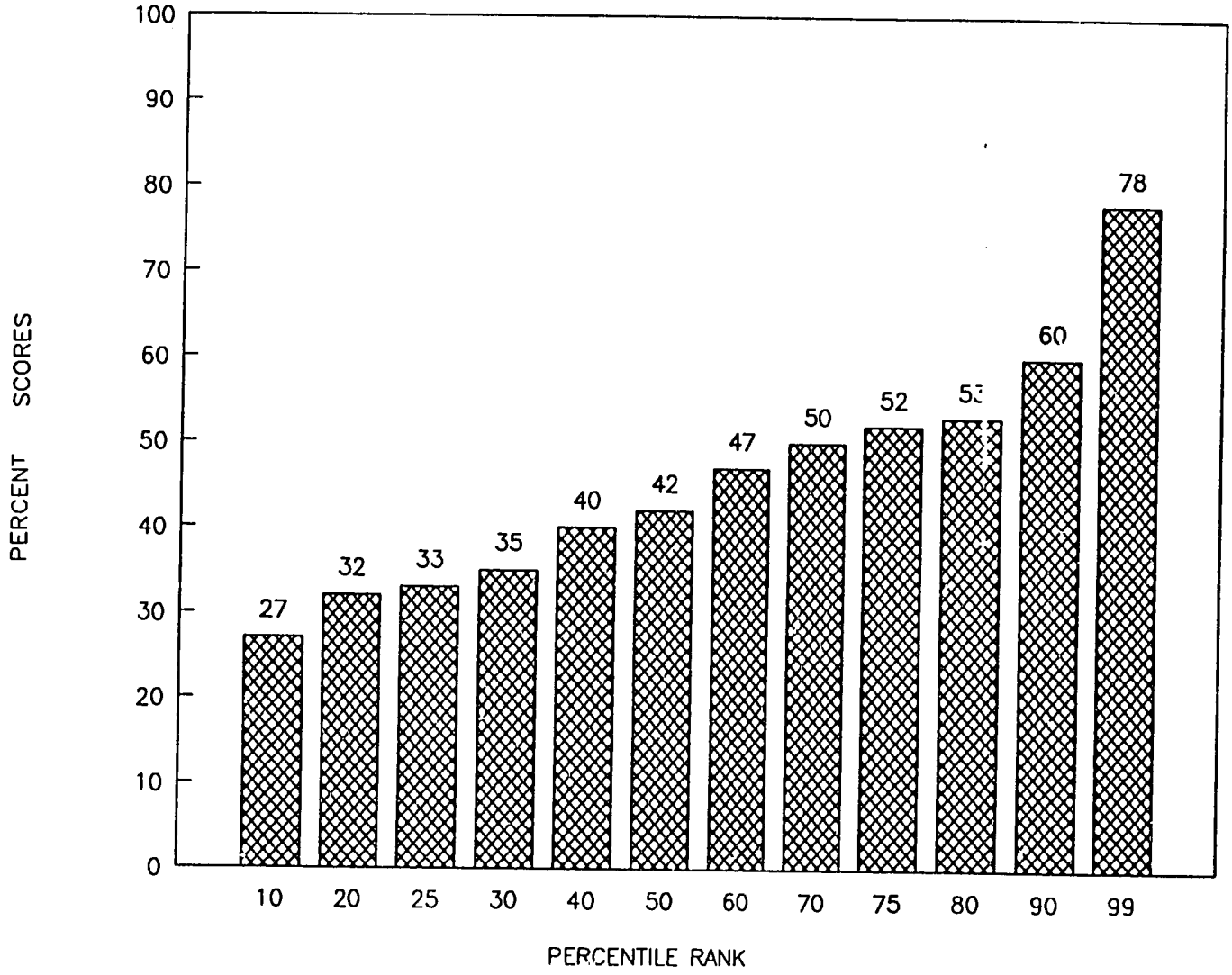


TABLE # 2

PERCENT SUCCESS-RATE ON SCIENCE ITEMS (K=60) BY PROVINCE BALOCHISTAN

ITEM NO/CONCEPT	P-VALUE / Biser		ITEM NO/CONCEPT	P-VALUE / Biser	
1. Animals	80	.00	26.Atoms & Molecules	57	.52
2. Animals	77	.20	27.Atoms & Molecules	45	.28
3. Animals	40	.23	28.Atoms & Molecules	51	.17
4. Animals	42	.26	29.Gases	36	.59
5. Animals	41	.40	30.Atoms & Molecules	52	.02
6. Animals	94	.57	31.Atoms & Molecules	46	.42
7. Animals	21	.32	32.Atoms & Molecules	71	.52
8. Animals	35	.05	33.Atoms & Molecules	50	.36
9. Insects	67	.12	34.Water	44	.30
10.Plants	50	.29	35.Atoms & Molecules	34	.45
11.Plants	43	.26	36.Water	79	.36
12.Plants	58	.45	37.Water	42	.43
13.Plants	99	.50	38.Water	27	.26
14.Plants	56	.32	39.Water	10	.03
15.Plants	44	.36	40.Water	13	.26
16.Plants	32	.20	41.Water	45	.35
17.Plants	67	.32	42.Force & Energy	18	.10
18.Plants	47	.43	43.Force & Energy	63	.34
19.Plants	65	.41	44.Force & Energy	24	.30
20.Foods	43	.48	45.Force & Energy	85	.51
21.Foods	56	.40	46.Force & Energy	23	.41
22.Foods	74	.20	47.Force & Energy	37	.22
23.Foods	45	.34	48.Gases	39	.49
24.Foods	18	.16	49.Gases	32	.12
25.Atoms & Molecules	37	.46	50.Light	54	.42

TABLE # 2 (Contd)

PERCENT SUCCESS-RATE ON SCIENCE ITEMS (K=60) BY PROVINCE BALOCHISTAN

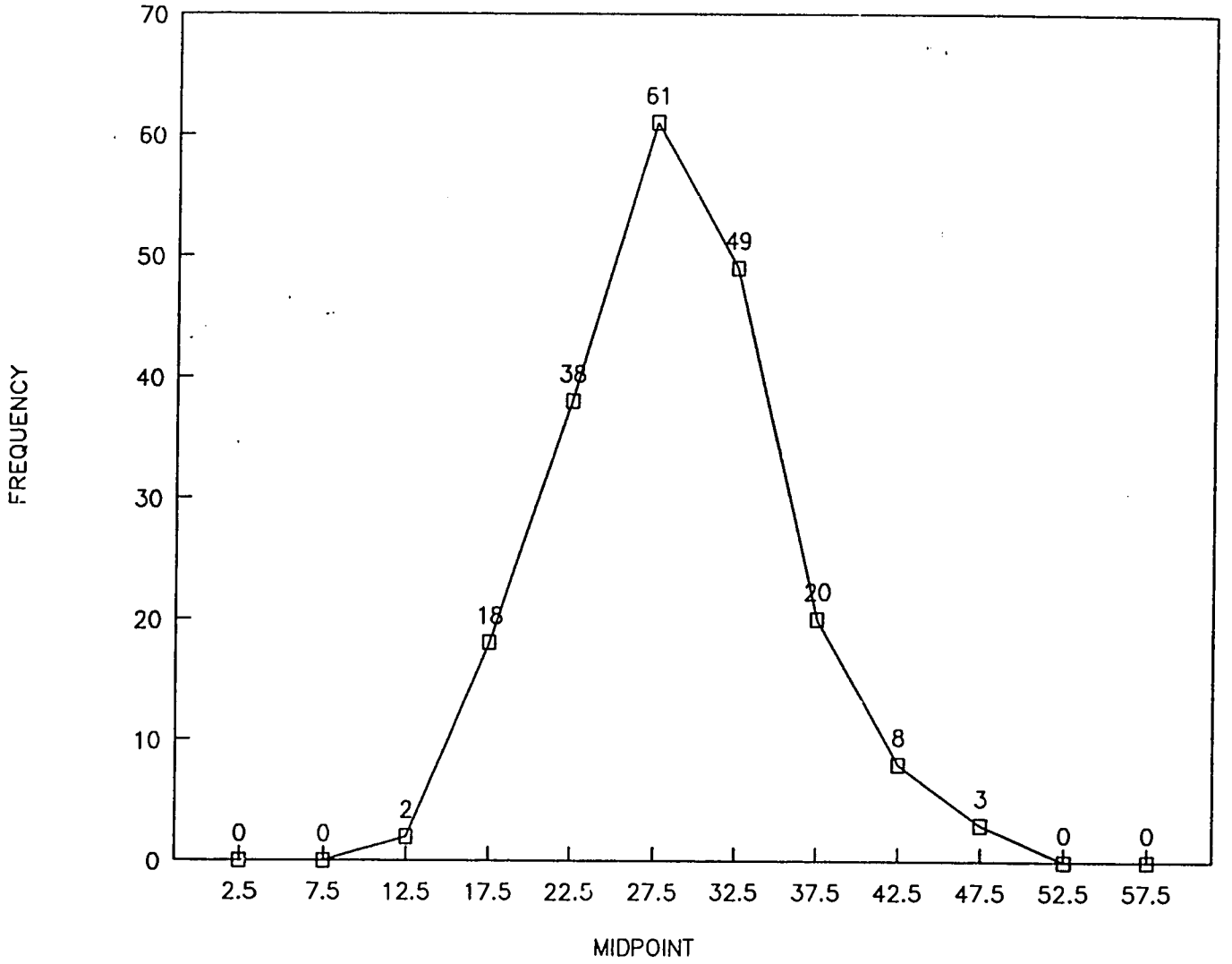
ITEM NO/CONCEPT	P-VALUE /Biser		ITEM NO/CONCEPT	P-VALUE /Biser	
51.Light	31	.30	56.Magnetism	32	.56
52.Light	46	.20	57.Magnetism	43	.12
53.Light	21	.27	58.Magnetism	40	.10
54.Magnetism	63	.39	59.Electro Static	16	.57
55.Magnetism	15	.25	60.Solar System	69	.46

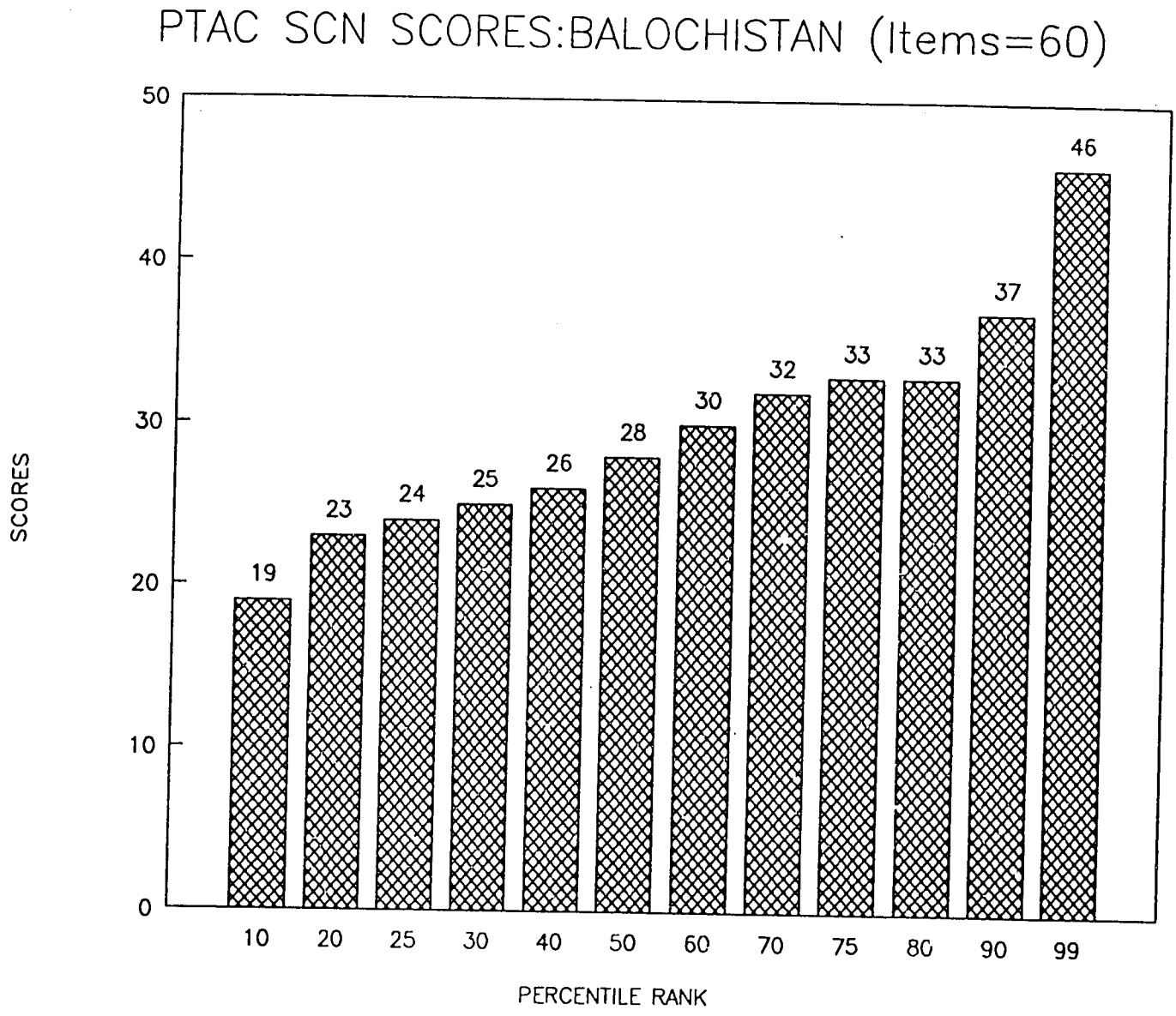
Mean=28.13, Percent=47, S.D= 6.53, n=199

Note: K indicates the total number of items in the test.

PTAC STUDENTS' SCN SCORES: BALOCHISTAN

FREQUENCY POLYGON




PTAC SCN SCORES:BALOCHISTAN (Items=60)
SCORES
50
40
30
20
10
0
19
23
24
25
26
28
30
32
33
33
37
46
10
20
25
30
40
50
60
70
75
80
90
99
PERCENTILE RANK

PTAC SCN SCORES:BALOCHISTAN (Items=60)

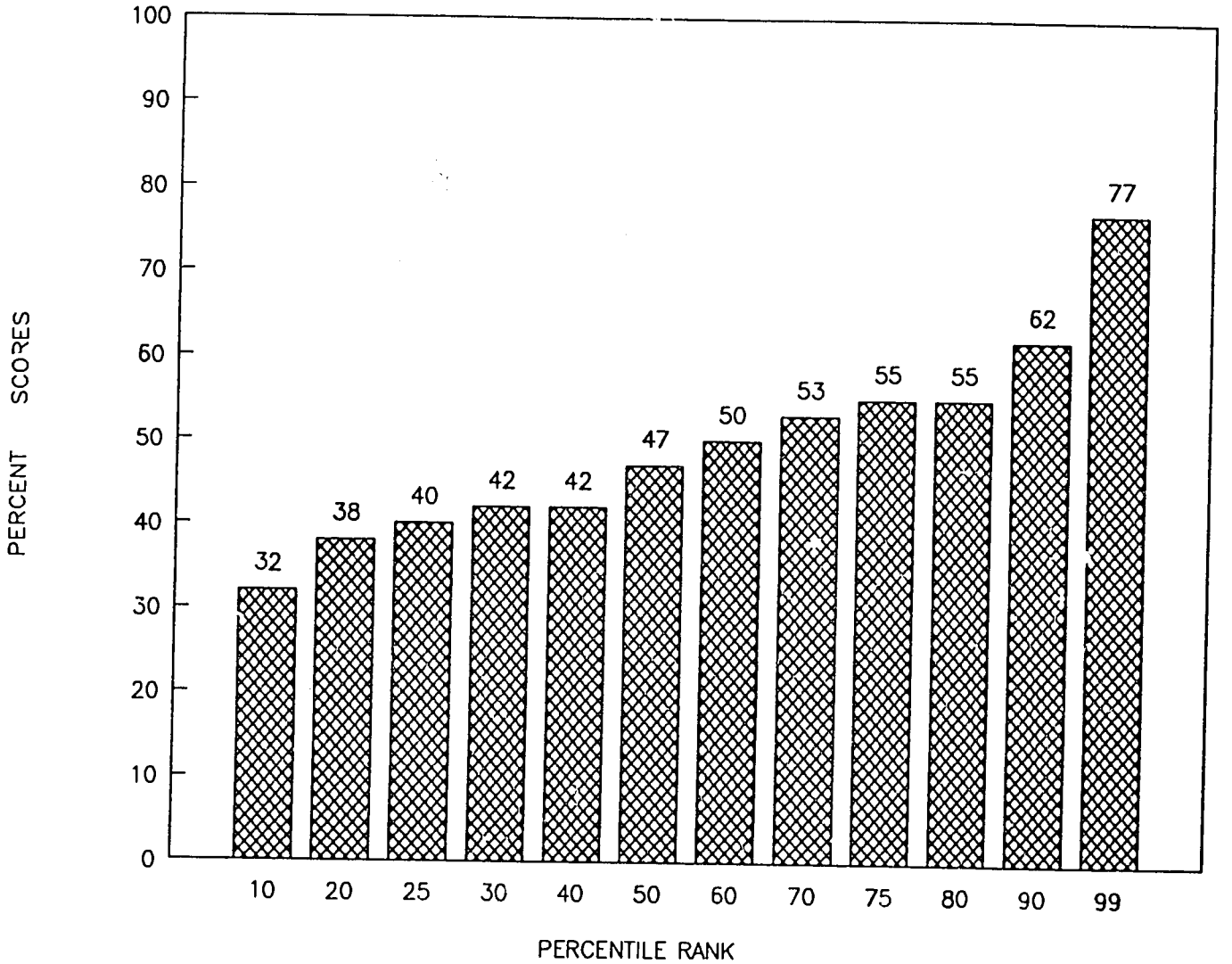


TABLE # 3
PERCENT SUCCESS-RATE ON URDU ITEMS (K=46) BY PROVINCE BALOCHISTAN

ITEM NO/CONCEPT	*P-VALUE/Bis*		*ITEM NO/CONCEPT*	*P-VALUE/Bis*	
1.Word Meaning:Interpretive Synonyms	81	0.34	14.Word Usage:Grammar	76	0.53
2.Word Meaning:Interpretive Synonyms	56	0.03	15.Word Usage:Grammar	83	0.70
3.Word Meaning:Interpretive Synonyms	73	0.59	16.Word Usage:Grammar	94	0.60
4.Word Meaning:Interpretive Synonyms	38	0.10	17.Word Usage:Grammar	70	0.45
5.Word Meaning:Interpretive Synonyms	35	0.50	18.Word Usage:Grammar	61	0.40
6.Word Meaning:Interpretive Synonyms	68	0.32	19.Word Usage:Grammar	36	0.36
7.Word Meaning:Interpretive Synonyms	30	0.48	20.Word Usage:Grammar	25	0.36
8.Word Meaning:Interpretive Antonyms	90	0.62	21.Word Usage:Grammar	37	0.46
9.Word Meaning:Interpretive Antonyms	44	0.57	22.Word Usage:Grammar	81	0.29
10.Word Meaning:Interpretive Antonyms	78	0.50	23.Word Usage:Spelling	17	0.60
11.Word Meaning:Interpretive Antonyms	76	0.50	24.Word Usage:Spelling	42	-0.03
12.Word Meaning:Interpretive Antonyms	36	0.26	25.Word Usage:Spelling	74	0.16
13.Word Usage:Grammar	79	0.50	26.Word Usage:Spelling	32	0.06

TABLE # 3(Contd)
PERCENT SUCCESS-RATE ON URDU ITEMS (K=46) BY PROVINCE BALOCHISTAN

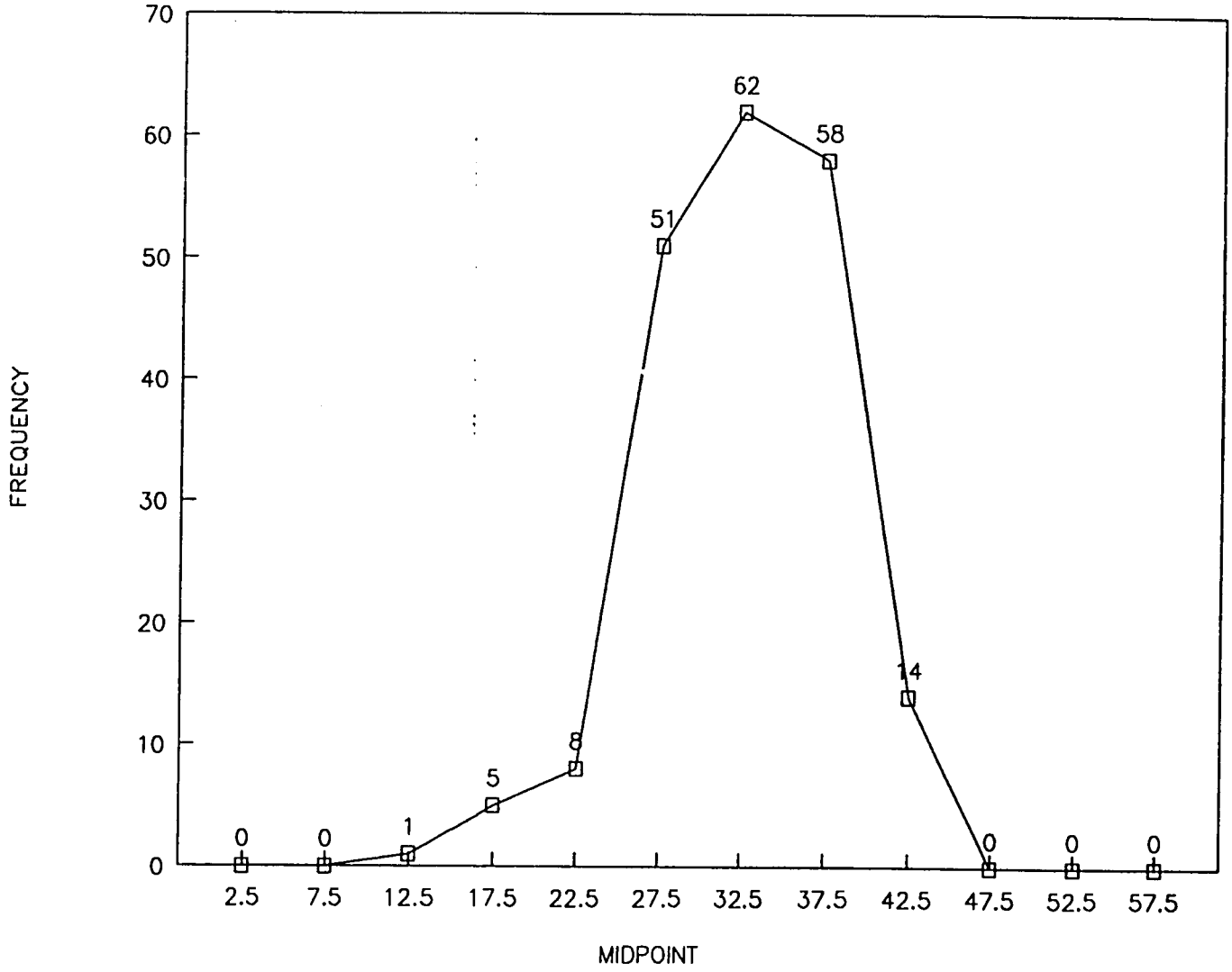
ITEM NO/CONCEPT	*P-VALUE/Bis*		*ITEM NO/CONCEPT*	*P-VALUE/Bis*	
27.Reading Comprehension (Title)	62	0.47	37.Sentence Completion (Grammar)	94	0.58
28.Deduction	88	0.51	38.Sentence Completion (Grammar)	82	0.47
29.Deduction	86	0.44	39.Sentence Completion (Grammar)	88	0.76
30.Reading Comprehension (Title)	32	0.38	40.Sentence Completion (Grammar)	82	0.65
31.Deduction	91	0.48	41.Sentence Completion (Grammar)	94	0.76
32.Deduction	78	0.57	42.Sentence Completion (Grammar)	93	0.71
33.Sentence Completion (Grammar)	98	0.62	43.Sentence Completion (Grammar)	86	0.41
34.Sentence Completion (Grammar)	96	0.91	44.Sentence Completion (Grammar)	72	0.61
35.Sentence Completion (Grammar)	92	0.86	45.Structure :Meaningful Order	78	0.54
36.Sentence Completion (Grammar)	68	0.51	46.Structure :Meaningful Order	84	0.54

Mean (Quetta)= 32.09, Percent = 70, S.D=5.39, n= 199

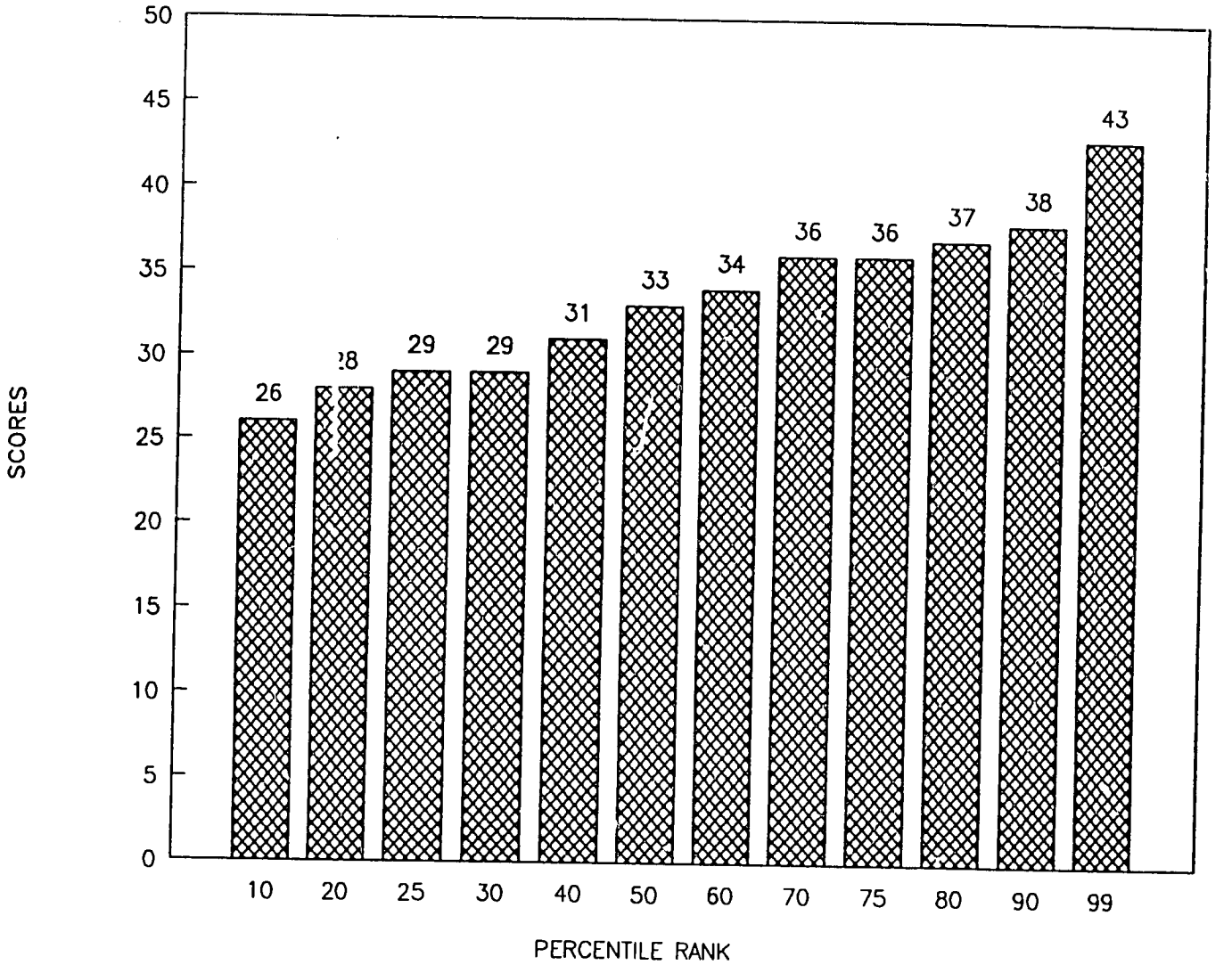
Note: K indicates the total number of items in the test.

PTAC STUDENTS' URDU SCORES: BALOCHISTAN

FREQUENCY POLYGON



PTAC URDU SCORES:BALOCHISTAN (Items=46)



PTAC URDU SCORES:BALOCHISTAN (Items=46)

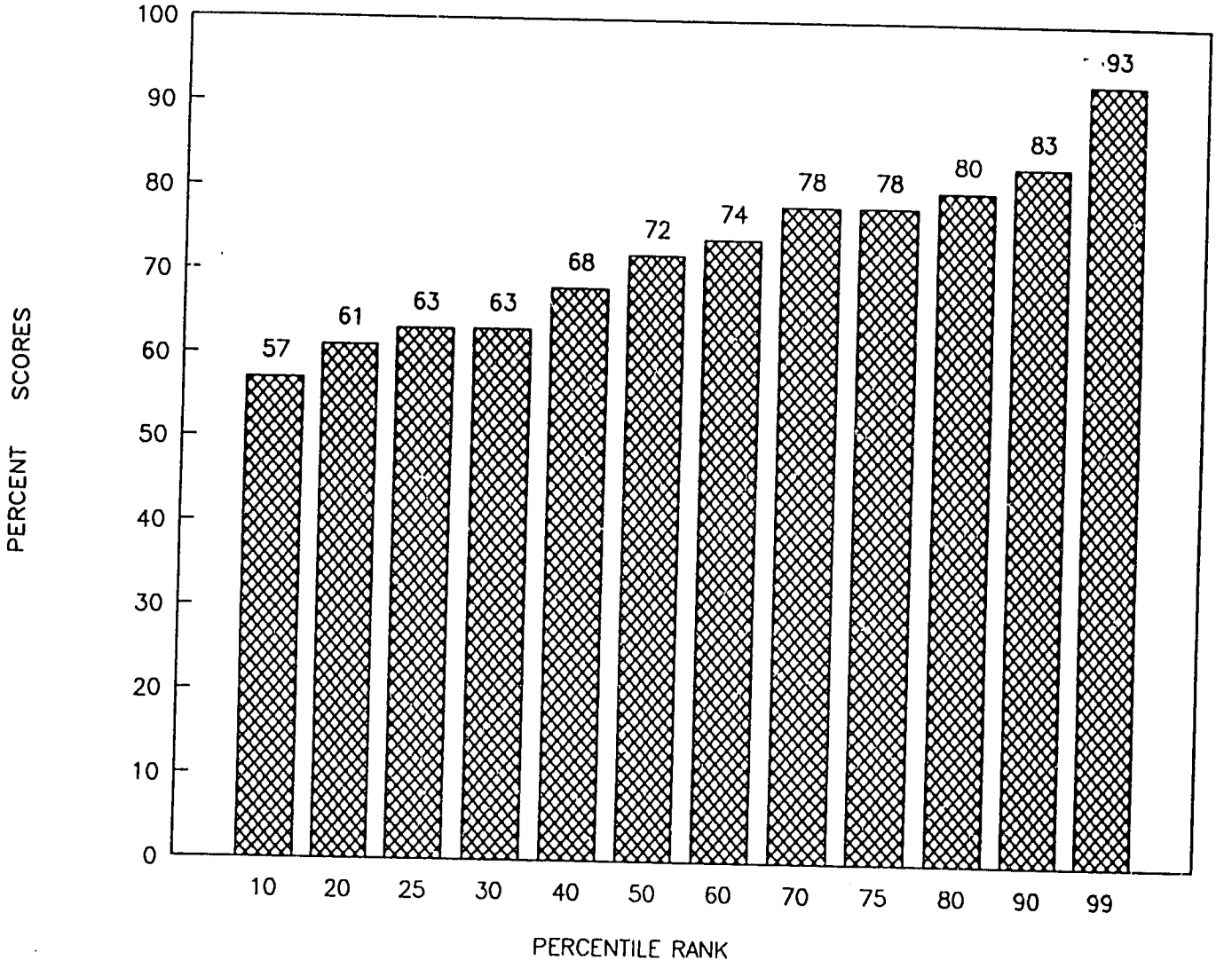


TABLE # 4

MATHEMATICS

Name of the Centre	Mean	%	S.D	n
G.H.S. Khudabadan Panjgur (M)	28.00	47	4.65	45
G.G.H.S. Khudabadan Panjgur (F)	20.70	35	6.82	27
G.G.H.S. Mastung (F)	19.00	32	6.90	28
G.College of E/Ed Mastung (M)	30.24	50	9.01	37
G.Special.S. Centre II Quetta (M)	26.63	44	7.91	30
G.G.H.S. Mission Road Centre II Quetta (F)	26.44	44	7.74	32

TABLE # 5

SCIENCE

Name of the Centre	Mean	%	S.D	n
G.H.S. Khudabadan Panjgur (M)	30.22	50	5.23	45
G.G.H.S. Khudabadan Panjgur (F)	24.33	41	5.25	27
G.G.H.S. Mastung (F)	27.68	46	7.47	28
G.College of E/Ed Mastung (M)	28.00	47	7.01	37
G.Special.S. Centre II Quetta (M)	26.93	45	7.28	30
G.G.H.S. Mission Road Centre II Quetta (F)	30.03	50	5.69	32

TABLE # 6

URDU

Name of the Centre	Mean	%	S.D	n
G.H.S. Khudabadan Panjgur (M)	32.02	70	3.87	45
G.G.H.S. Khudabadan Panjgur (F)	28.33	62	5.53	27
G.G.H.S. Mastung (F)	33.50	73	4.71	28
G.College of E/Ed Mastung (M)	31.24	68	5.45	37
G.Special.S. Centre II Quetta (M)	31.87	69	6.57	30
G.G.H.S. Mission Road Centre II Quetta (F)	35.31	77	4.43	32

TABLE # 7

Summary Statistics for Mathematics

Statistical Index	Value
No of Items	60
No of Students	199
Mean	25.70
Variance	65.41
Std.Dev	8.09
Skew	0.13
Kurtosis	-0.01
Minimum	6.00
Maximum	49.00
Median	25.00
Alpha	0.84
SEM	3.27
Mean p	0.43
Meam point-biserial	0.30
Mean biserial	0.41

TABLE # 8

Summary Statistics for Science

Statistical Index	Value
No of Items	60
No of Students	199
Mean	28.13
Variance	42.67
Std.Dev	6.53
Skew	0.27
Kurtosis	0.09
Minimum	13.00
Maximum	49.00
Median	28.00
Alpha	0.72
SEM	3.47
Mean p	0.47
Meam point-biserial	0.24
Mean biserial	0.32

TABLE # 9

Summary Statistics for Urdu

Statistical Index	Value
No of Items	46
No of Students	199
Mean	32.09
Variance	29.05
Std.Dev	5.39
Skew	-0.64
Kurtosis	0.77
Minimum	11.00
Maximum	43.00
Median	33.00
Alpha	0.76
SEM	2.62
Mean p	0.70
Meam point-biserial	0.31
Mean biserial	0.47

TABLE # 10

Defferences Between Subject Scores of PTAC Students

	Mean	S.D	diff	t	p
Urdu[1]	41.86	7 05			
Math	25.70	8 11	-16.15	-26.91	<0.001
Science	28.12	7 05	-13.73	-30.95	<0.001

[1] The actual number of items in this test was 46. For the purpose of this analysis, the Urdu test was brought to the same scale as mathematics and science test.

TABLE # 11

Analysis of Gender Differences for the Math Test

	Mean	%	S.D	n	F	p
Overall	25.70	43	8.10	199		
Male	28.37	47	7.30	112		
Gender					32.94	<0.0001
Female	22.26	37	7.83	87		

TABLE # 12

Analysis of Gender Differences for the Scn Test

	Mean	%	S.D	n	F	p
Overall	28.12	47	6.55	199		
Male	28.61	48	6.52	112		
Gender					1.22	.27 (NS)
Female	27.50	46	6.56	87		

TABLE # 13

Analysis of Gender Differences for the Urdu Test

	Mean	%	S.D	n	F	p
Overall	32.09	70	5.40	199		
Male	31.72	69	5.19	112		
Gender					1.18	.28 (NS)
Female	32.56	70	5.66	87		

APPENDIX K:
SUMMARY OF MEETING IN ISLAMABAD, MARCH 24,1994

Meeting at the Curriculum Wing
Federal Ministry of Education
March 24, 1994
11:30

Present: Mr. Abassi, Joint Secretary, Curriculum Wing, Federal Ministry of Education
Dr. Sar Biland Khan, AED (PED)
Dr. Kristian John Kirkwood AED (PED)

The purpose of this meeting was to review that status of the item development for the five subject areas, namely, English, biology, chemistry, physics, and mathematics.

Initially, the Dr. Abdul Wahab was supposed to be in attendance but was unable to attend the meeting for personal reasons.

However, Mr. Abbasi felt that it would be an opportune time to develop some strategies which could be utilized in accelerating the progress of the item development process.

After a lengthy general discussion about testing issues associated with Pakistan, the discussion focussed on how the task (the development of five item banks) could be completed as expeditiously as possible. According to Mr. Abassi, he felt that IBA should subcontract the development of the item banks to AED/PED since they had the expertise. Dr. Khan reminded him that he had no authority on this issue and Mr. Abassi should deal directly with Dr. Wade Robinson.

Mr. Abassi stated that he would discuss this possibility with Dr. Robinson in the immediate future.

The meeting was adjourned at 12:35 p.m.

APPENDIX L:
MINUTES OF MEETING WITH IBA, KARACHI, MARCH 31, 1994

Meeting at University of Karachi
March 31, 1994
11:00 A.M.

Present: Dr. Adbul Wahab, Director of IBA, University of Karachi
Mr. Abassi, Joint Secretary, Curriculum Wing, Federal Ministry of Education
Dr. Sar Biland Khan, AED (PED)
Dr. Kristian John Kirkwood AED (PED)

The meeting was started with opening remarks from Mr. Abassi. He stated that the purpose of the meeting was to get an update from IBA on the status of the development of the test items for the National Admission/Selection Test for the National Educational Testing Service (NETS). He also noted that he had been thinking about some alternatives that would enhance the development of a national item bank for English, mathematics, biology, chemistry, and physics. He had discussed the possibility of involving AED (PED). To this end he had discussed this with Dr. Wade Robinson, Chief of Party, AED (PED). He reported that Dr. Robinson is willing to get involved.

Mr. Wahab reported that there had been no developmental work done of the item bank. He distributed the summary sheet of the most recent examination for the Aga Khan Medical College. No psychometric characteristics were given for the this test. The national Board of Governors had not be established todate. This is to be done with an Act of Parliament.

A discussion ensued based on an article distributed by Dr. Khan. The article was printed in the newspaper DAWN which indicated that the Medical College were moving in the direction of a standard entrance examination. The discussion concluded with an agreement that interest in testing was a national trend and educators must be evolved and give direction to this movement.

The meeting then focused on the item bank development. Mr. Abassi then outlined how he thought the process of developing the item bank should evolve. There would be a working committee for each of the subject areas. The committees would be comprised of 4-5 individuals. He stated that one individual in each of the five subject areas would be identified as a

coordinator for the subject area. This individual would be an expert in the subject area in addition to having a national stature. The other members would include two subject specialists (one male, one female) who are currently teaching the specific subject. A test expert would also be a member of the committee. In addition, other vested interest group would be asked to contribute members to assist with the item development.

This process would involve approximately 25-30 individuals. Dr. Khan stated that this was a manageable group.

Mr. Abassi identified the various cities by subject area:

Karachi	Biology/Mathematics
Lahore	Physics
Peshawar	Chemistry
Islamabad	English/Mathematics

Mr. Abassi then stated that a 6 day workshop would be held in Peshawar during the last week of April (if possible). Green's hotel and other hotels such as the Shelton, government houses will be contacted to determine if they can host the participants. The actual meetings will take place at the PED offices. The alternative date would be sometime in May.

The participants would be paid TA and DA. A small honorarium would be given to each participants after the successful completion of their respective items. The funds will be distributed by IBA. It appears that there will be no subcontracting but rather IBA will play for any direct costs incurred. Mr. Absassi instructed Dr. Khan to outline the strategies required to implement this program and to have it ready before April 10,1994.

The process will be include the following steps:

1) IBA to contact Dr. Wade Robinson directly for a proposal including budget to be submitted to IBA.

2) A meeting of the major players will be held in Peshawar.

A tentative meeting will held in Islamabad on April, 1994. This will be confirmed by Mr. Abassi's office. The meeting was adjourned at 12:55 P.M.