DOCUMENT RESUME

ED 069 756 TM 002 211

AUTHOR Fritz, Kentner V.; Cornish, Richard D.

TITLE A User's Guide to Scoring and Improving Examinations Using the MERMAC Test Analysis and Questionnaire Package.

INSTITUTION Wisconsin Univ., Madison. Counseling Center.
PUB DATE Aug 71
NOTE 17p.
JOURNAL CIT Counseling Center Reports, v5 n3 August 1971

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS *Computer Programs; Evaluation; Higher Education; *Scoring; Statistical Analysis; *Test Construction; Test Interpretation; Test Results; *Tests
IDENTIFIERS *MERMAC Program

ABSTRACT

The MERMAC computer program is offered to the University of Wisconsin faculty for use in scoring and analyzing classroom tests. The characteristics of a good test are discussed; examples are given of the output of the MERMAC program; and the results are used to show how the quality of a test may be improved. Although the MERMAC Program is for scoring purposes only, it is suggested that a statistical analysis also be made to give necessary information for test improvement. Several options and levels of service are available. (Author/RS)

Counseling Center Reports

The University of Wisconsin 415 West Gilman Street Madison, Wisconsin 53706

Editor: Richard W. Johnson

Volume 5, Number 3 August, 1971

# A USER'S GUIDE TO SCORING AND IMPROVING EXAMINATIONS USING THE MERMAC TEST ANALYSIS AND QUESTIONNAIRE PACKAGE

Kenter V. Fritz and Richard D. Cornish

## Abstract

Many professors and/or their assistants spend a good deal of time both making and correcting tests. Considering the amount of time expended, it is desirable that a good test be developed. The MERMAC computer program is being offered to the faculty of the University of Wisconsin for use in scoring and analyzing classroom tests. The following pages briefly discuss the characteristics of a good test, give examples of the output of the MERMAC program, and explain how the results can be used to improve the quality of a test. Although the MERMAC program may be used for scoring purposes only, it is suggested that a statistical analysis also be made. This information not only tells how well a test measures what students have learned, but also gives the necessary information for test improvement. For the user's convenience, several options and levels of service are available.

# PREFACE

All of us engaged in the process of education are well aware of the turmoil over accountability. The public is rightfully asking us to prove that their tax dollar is being spent wisely. Such proof necessarily asks questions about what students have learned, the ability of the teacher and the success of the system in general. All of these variables, at one time or another, are measured by teacher-made tests. Periodic assessment of educational achievement is necessary to insure the quality of education. This assessment, to be valid, must be made with the best available instruments.

This document was designed to aid faculty in the preparation of good tests. We believe that it is non-technical enough to be used comfortably by those not familiar with statistics and test construction, yet detailed enough to aid the more sophisticated. Because of space limitations the document also deals with only the technical aspects of test construction. For a more detailed discussion of these and other aspects of test construction the reader is referred to: Ebel, Robert L. Measuring Educational Achievement, Englewood Cliffs, N. J.: Prentice-Hall, 1965. Also, the authors are available to consult with faculty members having specific problems.

K.V.F.
R.D.C.

## VALIDITY[1]

One important characteristic of a good test is its validity--does it measure what it is supposed to measure? Normally the instructor making a test thinks of questions and decides whether or not they bear a relationship to the class material, particularly the objectives of the class. These judgments by a single individual are subject to a large margin of error. To be certain of validity, and to improve the quality of a test, each test item should be judged by a number of qualified individuals. Under the teaching assistantship framework at the University of Wisconsin, this can be accomplished by having an assistant devise test items, then giving them to other assistants for criticism and finally referring them to the professor for approval.

When agreement is reached that items are valid, they should be included in the test, provided they have met, or can be reworded to meet, certain other good measurement characteristics discussed below.

## RELIABILITY[1]

The type of reliability measured by MERMAC refers to the consistency of a test. For example, if the test were split into two halves, the students' score on the first half should be similar to his score on the second half. Reliability may vary from 0.00 to 1.00. As the reliability approaches 1.00, the test scores reflect a higher degree of consistency. A good classroom test should have a reliability coeffient of at least 0.80, preferably 0.90. Reliabilities greater then 0.90 are difficult to obtain without making the test prohibitively long.

[1] Both Validity & Reliability, as discussed herein, are limited in scope. Interested persons are referred to Ebel or the authors for further information.

A reliability of less than 0.80 is an indication that a test needs revision. Methods of improving test reliability are discussed below.

## STANDARD ERROR OF MEASUREMENT

The standard error of measurement (SEM) is an extremely important characteristic of tests, yet one that is often overlooked. It is closely related to reliability and fluctuates with it. The SEM indicates how much error is inherent in a measurement technique. Obviously we would not place as much faith in the weight of a small object when using a bathroom scale as we would if weighed on a precision scale from a chemist's laboratory.

Likewise, measurements obtained by tests always have some degree of error in them. Consider a raw score of 50. If we know that the test has a SEM of 1 raw score point, we know that the true score is likely to be somewhere between 48 and 52. (It is customary to consider the limits as 2 times the SEM above and below a particular score.) The greater the SEM, the poorer the reliability and therefore the greater the range of error.

The relevance of this statistic to good measurement can be seen if we consider a test with a cutoff point of 35. If the SEM is 1 it is unrealistic to say that a person who obtained 34 failed while a person who obtained 36 passed. A small SEM tells us their scores are very similar, but it is still possible that the person who scored 34 would have a true score higher than the person who scored 36. Tests with low reliability and consequently a large SEM should not be used to discriminate between persons falling on or near the cut off point.

In general, two things may be done to improve the measurement characteristics of a test, (1) add more items, and (2) improve the quality of items, e.g., increase the correlation among the items. Both of these methods are discussed in more detail below.

## Length of Test

In order for a test to be a good instrument it should usually be at least 75 to 100 items long. A test of fewer items generally will not have suitable measurement characteristics.

lf you have a test that does not contain the desired characteristics and you think this can be improved by lengthening the test, the Spearman-Brown Prophecy Formula, available in the MERMAC package, will tell you how much longer the test has to be for the reliability to be satisfactory. If the test contains about 100-150 items and still does not have the desirable measurement characteristics, it is usually preferable to improve the quality of the items rather than to increase the length of the test.

## Quality of Items

The quality of items is improved by using the information provided by an item analysis. Item analyses provided by the statistical program called MERMAC consist of two main parts: (1) a table of responses by fifths, and (2) a graphic presentation of correct responses by fifths. This procedure is explained more fully below.

Item Differentiation. If a test is properly differentiating between students who do well on the test and those who do not, then a greater number of students in the first fifth should choose the correct response than students in the lowest fifth. In other words, students who do well on the total test should be expected to get a particular item correct more frequently than those students who do poorly. If there is no difference between the groups, then the item is not differentiating among the groups and is therefore not useful in terms of assigning class rank or grades.

Distractors. Distractors are choices inserted in a multiple choice question as alternatives to the correct response. Those that nobody chooses are "dead wood" and really add nothing to the test. Distractors should be plausable enough so that they will be attractive to those who do not know the material fully. Distractors should be responded to more often by students in the lower fifths than by students in the upper fifth. When distractors are poor, students can arrive at the correct answer by a process of eliminating the obviously wrong answers even without a good command of the subject matter.

Item Difficulty. On a test that incorporates ideal measurement practices, 50 percent of the students should get 50 percent of the test correct. This means that on a workable basis the proportion of correct responses to most of the test items should be from .4 to .6. The MERMAC item analysis yields a proportion (PROP) of people choosing each alternative and should be checked to determine the difficulty of an item. Items that are too easy can be increased in difficulty by making the distractors more plausible.

Point=Biserial Correlation Coefficient. The point-biserial correlation coefficient (RPBI) indicates how much predictive power an item has. It serves further to indicate the effectiveness of the item and is a very important statistic in item analysis. The possible range of RPBI is from -1.0 to +1.0 with a score of +0.20 being generally the lowest acceptable value for the correct response and a value of +0.40 or above considered good. If the RPBI value is lower than +0.20, the item as written should not be used in the test. Changes in the item can be made, however, and the test statistics will help locate the trouble. For example, check the question for possible ambiguity. Also check to see if an item is too easy or too difficult. Use the item analysis to see whether or not the item is differentiating among the good and poor students, and examine the distractor items to see that they are working properly.

Figure 1 is a sample of the most elementary analysis available from MERMAC. For each person a raw score, standard score, and percentile is given. The standard scores and percentiles are derived from local norms, that is to say, based on the test as it was actually given. Particular attention should be paid to the standard score since it serves as a means of comparing scores on different tests. The standard score from one test is directly comparable to the standard score of another even though the ranges of attainable scores may be different. In Figure 1 the standard scores have a mean of 500 and a standard deviation of 100.

Figure 2 is an example of a slightly more sophisticated output in that it not only gives the test scores, but their distribution. Note that the obtained raw scores range from 4 to 11. The total number of people who obtained each score is listed in the frequency table and is presented graphically in the test frequency distribution. The standard scores corresponding to each raw score are listed so that comparisions may be made from one test to another. In addition, percentiles, percentages and cumulative frequencies are given.

Figure 1

*** MERMAC -- TEST ANALYSIS AND QUESTIONAIRE PACKAGE ***

TEST ANALYSIS FOR DR. JOHN DOE OF EDUCATION 100

| | | DISCUSSION | | | READING | | | LECTURE | | | TOTAL | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| NAME | | RAW | STAND | PCT | RAW | STAND | PCT | RAW | STAND | PCT | RAW | STAND | PCT |
| ABBOTT | JUDITH | 11 | 610 | 99 | 8 | 447 | 37 | 25 | 613 | 92 | 44 | 592 | 88 |
| ABERLE | JOHN | 11 | 610 | 99 | 9 | 514 | 67 | 23 | 545 | 70 | 43 | 569 | 79 |
| ABRAMSON | BARB | 10 | 540 | 78 | 11 | 649 | 99 | 25 | 613 | 92 | 46 | 636 | 96 |
| ADAMS | PAMELA | 10 | 540 | 78 | 9 | 514 | 67 | 18 | 375 | 15 | 37 | 435 | 26 |
| ADELMAN | KATHY | 11 | 610 | 99 | 10 | 582 | 88 | 22 | 511 | 58 | 43 | 569 | 79 |
| ADLER | JOEL | 10 | 540 | 78 | 11 | 649 | 99 | 23 | 545 | 70 | 44 | 592 | 88 |
| ALLAR | MATTHEW | 9 | 470 | 41 | 9 | 514 | 67 | 20 | 443 | 31 | 38 | 458 | 33 |
| BAEDER | RAY | 10 | 540 | 78 | 9 | 514 | 67 | 21 | 477 | 42 | 40 | 502 | 49 |
| BAILEY | THOMAS | 10 | 540 | 78 | 11 | 649 | 99 | 24 | 579 | 82 | 45 | 614 | 92 |
| BARRY | JAMES | 10 | 540 | 78 | 9 | 514 | 67 | 26 | 647 | 98 | 45 | 614 | 92 |
| BAUMANN | DONALD | 8 | 400 | 20 | 7 | 379 | 19 | 15 | 273 | 3 | 30 | 279 | 4 |
| BECK | GAIL | 9 | 470 | 41 | 7 | 379 | 19 | 21 | 477 | 42 | 37 | 435 | 26 |
| BECKER | PAMELA | 9 | 470 | 41 | 9 | 514 | 67 | 23 | 545 | 70 | 41 | 525 | 59 |
| BENTON | OTTIE | 10 | 540 | 78 | 10 | 582 | 88 | 22 | 511 | 58 | 42 | 547 | 68 |
| BERNSTEIN | IRIS | 10 | 540 | 78 | 9 | 514 | 67 | 23 | 545 | 70 | 42 | 547 | 68 |
| CARLSON | PAT | 10 | 540 | 78 | 9 | 514 | 67 | 23 | 545 | 70 | 42 | 547 | 68 |
| COLLIER | DANIELE | 11 | 610 | 99 | 5 | 244 | 3 | 18 | 375 | 15 | 34 | 368 | 12 |

Figure 2

```
                 *** MERMAC -- TEST ANALYSIS AND QUESTIONAIRE PACKAGE ***

                     TEST ANALYSIS FOR DR. JOHN DOE OF EDUCATION 100

                                         LECTURE

                                                                        TEST FREQUENCY DISTRIBUTION
RAW                STANDARD    PER-                                CUM
SCORE               SCORE    CENTILE      PERCENT      FREQ.    FREQ.
                                                                          EACH * REPRESENTS 3  PERSONS
 11                  649        99          12.1         42       346   **************
 10                  582        88          20.8         72       304   *****************
  9                  514        67          30.3        105       232   ***********************************
  8                  447        37          17.6         61       127   ********************
  7                  379        19          12.1         42        66   **************
  6                  312         7           4.0         14        24   *****
  5                  244         3           2.3          8        10   ***
  4                  177         1           0.6          2         2   *
```

If you were to read the table across columns for a raw score of 6, you would have the following information. The standard score is 312. The percentile is 7 meaning that approximately 7% of the students obtained a score of 6 or less. Fourteen students (from the frequency column) or 4.0 percent of those taking the test obtained a score of 6. From the cumulative frequency column you can see that 24 people had raw scores of 6 or less.

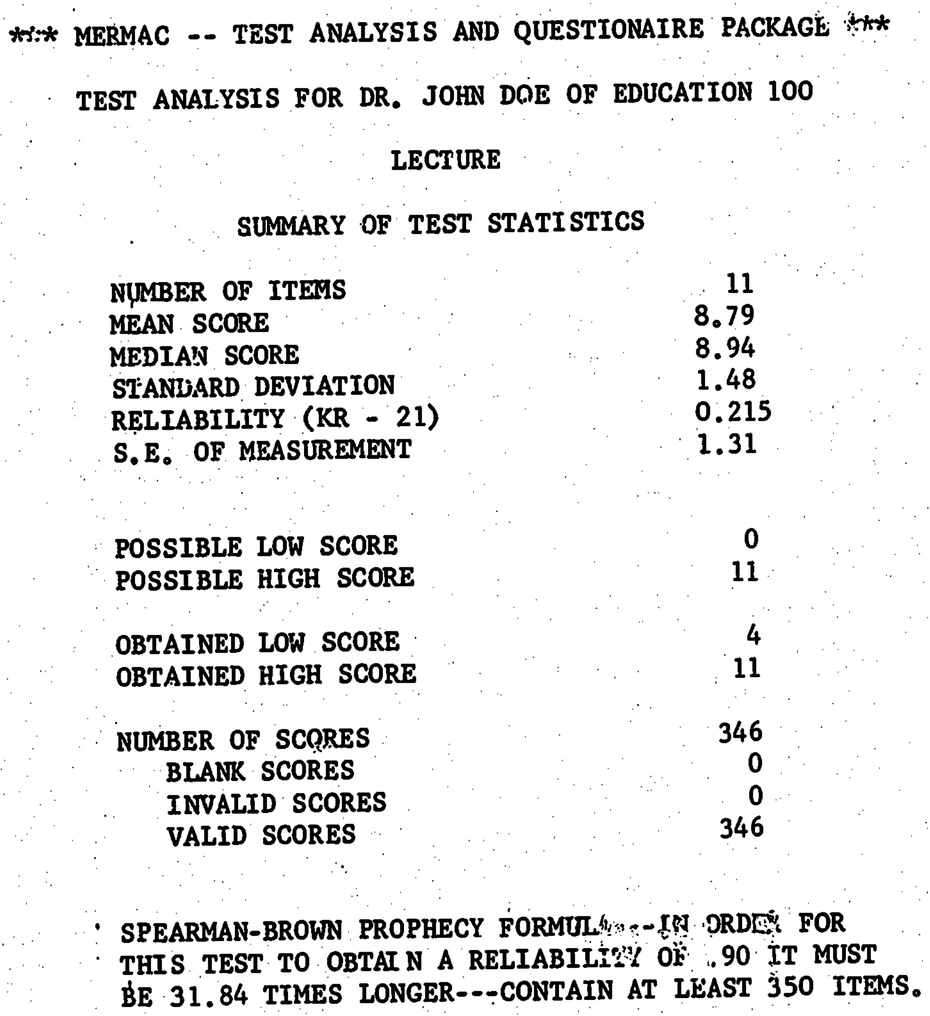
Figure 3 shows a summary of test statistics for a test having a possible score range of 0 to 11. The lowest score actually obtained was 4 and the highest 11. The standard error of measurement of 1.31 means that for any score obtained on the test, the true score actually falls within a plus or minus 2.62 of that score. For example, an obtained score of 7 represents a true score of somewhere from 4.38 to 9.62.

The standard error of measurement is inordinately large for this size test and consequently the reliability is low. The Spearman-Brown Prophecy Formula states that in order for the reliability of this test to be .90, it must contain at least 350 items of equal characteristics to the 11 already in the test. It would be wise, in cases where the reliability is very low, to check the items and improve or replace poor ones.

Other important information to check is the number of valid scores. If there are a large number of blank or invalid scores, the test statistics will not be accurate. As a matter of interest, the close values of the mean and median indicates that the distribution of test scores is fairly well balanced.

Figure 4 shows an example of an item analysis. The class is divided as closely as possible into fifths (quintiles) according to performance

Figure 3



```
*** MERMAC -- TEST ANALYSIS AND QUESTIONAIRE PACKAGE ***

    TEST ANALYSIS FOR DR. JOHN DOE OF EDUCATION 100

                        LECTURE

               SUMMARY OF TEST STATISTICS

      NUMBER OF ITEMS                         11
      MEAN SCORE                            8.79
      MEDIAN SCORE                          8.94
      STANDARD DEVIATION                    1.48
      RELIABILITY (KR - 21)                 0.215
      S.E. OF MEASUREMENT                   1.31


      POSSIBLE LOW SCORE                       0
      POSSIBLE HIGH SCORE                     11

      OBTAINED LOW SCORE                       4
      OBTAINED HIGH SCORE                     11

      NUMBER OF SCORES                       346
         BLANK SCORES                          0
         INVALID SCORES                        0
         VALID SCORES                        346


      SPEARMAN-BROWN PROPHECY FORMULA---IN ORDER FOR
      THIS TEST TO OBTAIN A RELIABILITY OF .90 IT MUST
      BE 31.84 TIMES LONGER---CONTAIN AT LEAST 350 ITEMS.
```

Figure 4

```
                                        A IS CORRECT RESPONSE
ITEM 1  Percent of correct responses    MATRIX OF RESPONSES BY FIFTHS
                                                 A     B     C     D    E  OMIT
1ST                            *        1ST     68     1     2     0    0    0
2ND                          *          2ND     66     1     4     1    0    0
3RD                      *              3RD     72     1    11     5    0    0
4TH                    *                4TH     35     1     9     2    0    0
5TH             *                       5TH     34     4    15    14    0    0
   10 20 30 40 50 60 70 80 90 100       PROP  0.79  0.02  0.12  0.06 0.00 0.00
                                        RPBI  0.41 -0.10 -0.24 -0.30 0.00 0.00
```

on the total test. Then each item is broken down according to the way each fifth answered it. Thus, 68 persons in the first fifth (of the total test) chose alternative A (the correct response), 1 person chose B, 2 people chose C, and none chose D or E.

Since the PROP for the correct answer is 0.79, this would be a relatively good item - the RPBI being +0.41. Distractors D and E are not working as they should and the item could be improved by making them more plausible. Making the distractors more plausible will make the item more difficult, and lower both PROP and RPBI. Care should be taken to insure that these stay within acceptable limits.

Figure 5 illustrates a "bad" item and a quick glance at the graph will make this clear. Note that the responses are all clustered at the far right end. This is an indication that the item is making no discrimination among students, and since the responses are all clustered at the right of the graph the question is obviously too easy. The table of responses will bear out this information. The number of correct responses (choice A) is nearly the same for students in the top fifth (70) as it is for those in the bottom fifth (67). The PROP is 0.98, or in other words, 98 percent of the students chose the correct response. Distractors D and E are not functioning at all since no students chose them. The RPBI for this item is only -0.07, a value far below tha minimum acceptable value of +0.20.

It may be pointed out that a few "easy" items as shown above may be retained as "stimulators" so that students will not consider the test too difficult. Also,certain basic material may have been emphasized to the point where every student is expected to know it. However, the fact still remains that such

Figure 5

```
ITEM 2  Percent of correct responses

1ST                                    *
2ND                                  *
3RD                                   *
4TH                                   *
5TH_____________________________________*
    10 20 30 40 50 60 70 80 90 100
```

A IS CORRECT RESPONSE

MATRIX OF RESPONSES BY FIFTHS

|  | A | B | C | D | E | OMIT |
|---|---|---|---|---|---|---|
| 1ST | 70 | 1 | 0 | 0 | 0 |  |
| 2ND | 68 | 2 | 2 | 0 | 0 |  |
| 3RD | 87 | 1 | 1 | 0 | 0 |  |
| 4TH | 46 | 0 | 1 | 0 | 0 |  |
| 5TH | 67 | 0 | 0 | 0 | 0 |  |
| PROP | 0.98 | 0.01 | 0.01 | 0.00 | 0.00 |  |
| RPBI | -0.07 | -0.80 | -0.01 | 0.00 | 0.00 |  |

items do not contribute to grade differentiation and the burden must be borne by those remaining items which do possess good discriminating ability.

* * *

The Counseling Center, as a service unit of the University of Wisconsin, is prepared to aid faculty members who may wish to initiate measurement procedures with respect to classroom testing. This can vary anywhere from simple test scoring, a decided time-saver in the case of large classes, to a more complete statistical analysis as outlined in this report.

Faculty members are invited to contact the Counseling Center for details.

415 West Gilman Street
Phone: 262-1744

Terms with which you may not be familiar have been listed below giving definitions and explanations of their usage for you convenience.

## GLOSSARY

Cumulative frequency - the number of times a particular score (or a lower score) is obtained. It is found by adding the frequency of a particular score to the frequency of all the scores below it.

Frequency - the number of times which a particular event occurs, or for purposes of test analysis, the number of students who obtained a particular score.

Mean score - the average of all the scores or the total value of the scores divided by the number of scores.

Median score - the point on the score scale which has 50% of the scores above it and 50% of the scores below it. For example, the set of scores 4, 4, 7, 9, 10,would have a median score of 7 since there are 2 scores above 7 and 2 below. In an even numbered series, the median would be half-way between the two middle numbers. If 7 were missing, the above series would have a median of 6.5.

Percentile - another relative position measure which will indicate a person's position within a group. For example, if a student is placed at the 60th percentile, assuming everyone took the same test, he did better than 60% of the students, and not as well as 40% of the students. Another useful interpretation, involving a slight approximation, is to consider the percentile as a cross-section of a group (100 units) with 50 at the half-way point, 25 the one-fourth, 75 the three quarters, and any other score at some point on this scale of 100. It is even useful to think of each person as standing in a line of 100 persons.

Point-biserial correlation coefficient (RPBI) - point-beserial correlation may range from a value of -1.0 to +1.0. If a test item is a "good" predictor, it will have an RPBI of not less than +0.20 and preferably a value of +0.40 or better. Checking for anbiguity in the question or in the response choices, difficulty of the question, effectiveness of distractor items, etc., will enable you to locate the source of difficulty within the test item.

Raw score - actual score of the student stated as the number of items corre t. Sometimes a formula to correct for guessing is used.

Reliability - reliability is an indication of how consistently a test measures. It may vary from 0.00 to 1.00 but the closer the reliability is to 1.00, the more stable the test scores are, i.e., a small degree of chance fluctuations. For classroom purposes the reliability should be at least 0.80 and preferably above 0.90. A reliability of less than 0.80 is an indication that there is a need for revision. You may improve your test by improving the individual items (check your item analysis) or by adding items to the test (check your item analysis) or by adding items to the test (check Spearman-Brown prophecy formula). There are various reliability formulas. A common one is the Kuder-Richardson number 21 (KR21).

Standard error of measurement - closely related to reliability, this statistic will tell you the amount of measurement error to be found in your test. A low reliability, and consequently a high standard error of measurement, will mean several things in terms of your test one of which is that you cannot accurately assign grades for scores around the cutting off points.

Spearman-Brown prophecy formula - this formula will predict the number of additional items required to bring your test to an acceptable level of reliability.

Standard deviation - the amount of scatter in a distribution of test scores. A large value means wide scatter and a small value means a "compact" distribution with small deviation about the mean. Suppose a certain test had a mean of 35 and a S.D. of 4, we would find, roughly, two-thirds of the scores between 31 and 39. Also, we would expect only occasionally a score below 23 or above 47, calculated by taking 3 times the S.D. and subtracting from or adding to the mean. If a distribution is skewed, i.e., "bunched" toward one end or the other, then the above interpretations are less accurate.

Standard score - a relative position measure by means of which scores in two different tests may be compared. For example, a score of 35 on a 40 item test may be comparable to a score of 70 on an 80 item test. However, given just the scores 35 and 70, it would be difficult to tell whether the scores were really comparable. Use of standard scores will enable you to determine that one score is actually better than another. This could not be known from raw scores alone. For example, 35 could actually be better than 70, if 35 has a higher rank or position in its distribution than 70 has in the other distribution.