ABSTRACT

A Test Response System was developed to display and analyze student results on multiple response tests using optically scanned examination forms. TRS combines latest developments in optical scanning technology with SAS® 6.04 and SAS® 6.06. A weighting system allows examiners to omit some questions and emphasize others.

Files created with optical readers are saved as SAS® files with SAS® 6.04 and transmitted with SAS® CONNECT for processing with SAS® 6.06.

The output shows the student’s total correct, total wrong, omitted, and multiple responses; the student’s weighted total; adjusted weighted total according to five levels of guessing penalties; and the student’s class rank.

For each item the report includes the frequency of each answer (including response and multiple responses); item variance; and item-test correlation. Psychometric measures include Kuder Richardson 20 index of reliability (Kuder and Richardson, 1937) and coefficient alpha (Cronbach, 1951).

TRS incorporates its own monitoring system so that examiners may compare courses and sections and graph grades historically.

INTRODUCTION

Psychometric analysis examines the extent to which a test achieves its primary purpose of differentiating reliably among individuals. Test quality depends upon the test maker and the individual test maker may be very different. Thus the first requirement is that responses to an item vary. If everyone gives the same answer to a question the item is clearly insensitive to varying skill levels among the respondents.

The second requirement is that item variances are meaningfully related to the construct being measured. In many testing situations, an operational definition of the construct is the total test score. Thus the higher the item-test correlation, the more the item appears to be measuring what it is intended to measure.

Provision of item statistics facilitates the process of refining a test for reuse. A more immediate problem is interpretation of the current test score. Since this is a composite, its meaning will be most clear when the contributing items form a homogeneous set. An item pool is homogeneous when only a small proportion of each item variance is specific to that component; that is, item covariances are large. A test composed of such items is internally consistent. Split-half methods of assessing consistency divide a test into two pieces and correlate the halves. This approach is flawed, however, since it assumes the items are grouped together and it is therefore not unique.

Coefficient alpha is an index of internal consistency based on the variance of the test scores and the covariances of the contributing items. Alpha is the mean of all possible split-half reliability coefficients and provides a lower bound estimate of the proportion of test variance attributable to common factors among the items. Cuciuba, Kenny and Costelloe (1980) have examined the distribution of coefficient alpha in a Monte Carlo simulation study and provided routines to test the probability that an obtained alpha is significantly different from zero (no consistency). Sampling theory for coefficient alpha is discussed by Feldt, Woodruff and Salih (1987).

Apart from interest in the characteristics of the test, there is often concern with the reliability of the assessments made of the individual. If a test subject’s true score is defined as the average of scores that would be obtained over an infinite number of repeated testings with the same test, the score obtained on any one occasion may be considered a random selection from this hypothetical distribution. The obtained score, therefore, is determined in part by that individual’s true standing on the construct measured and in part by sampling error. The psychometrist is interested in assessing the probable margin of error and in estimating true scores.

Application of psychometric theory with SAS®

Ryerson Polytechnical Institute provides examination results using SAS® 6.06 in a VM system with CMS. The software, termed Test Response System, is a practical application of the psychometric principles described in the Introduction. TRS has been used to score thousands of examinations since its inception, and it is a successful example of how SAS® can improve efficiency in a large educational institution.

TABULAR REPORT

Questions and responses are tabulated with PROC FREQ. This provides the data for all subsequent psychometric results.

Example 1 shows that the number of students is constant for all items. Column totals show the number of times an answer is chosen. The TITLE command uses a macro to display the course and section designations. Frequencies show the responses, including nonresponse and multi-response, for each item. Results reveal how the wording or order of the questions affects responses, and column percentiles show whether there is a concentration of nonresponse or multi-response answers in any question.

STATISTICAL REPORT

Psychometric measures are shown in Example 2.

Item difficulty

When the responses are scored as either right or wrong, item difficulty (p) is the proportion of the sample answering the item correctly. Paradoxically, therefore, the higher the value the easier the item. Values range from .00 (no-one answered correctly) to 1.00 (all respondents answered correctly). Statistically, the closer the value is to .5 the more discriminating the item.

Item variance

Item variance is the product of the proportion of respondents who select that question (p) and the proportion who fail to solve it (1-p). Thus if item difficulty is .60, the variance for that item is .60 x .40 = .24. Values range from .00 to .25. The higher the value the greater the potential contribution of the item to test reliability. Maximum differentiation is achieved when p = .5 and the item variance is .25. However, since the obtained item difficulties tend to be inflated by guessing, true p values are likely to be lower. Therefore, it has been suggested that items with observed p values higher than .5 might be optimal (Lord, 1952).

Item test correlation

For dichotomous scoring, the item test correlation is a point biserial coefficient which indicates the extent to which an item measures the same attributes as the test as a whole. Values range between -1.00 and +1.00. Positive values indicate that individuals
### Example 1

**Table of Question by Response**

<table>
<thead>
<tr>
<th>Question</th>
<th>Response</th>
<th>Multi-</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.05</td>
<td>0.00</td>
<td>12.36</td>
<td>23.56</td>
<td>15.38</td>
<td>13.39</td>
<td>23.00</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.00</td>
<td>0.00</td>
<td>13.46</td>
<td>13.56</td>
<td>23.46</td>
<td>13.56</td>
<td>12.38</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>13.36</td>
<td>12.38</td>
<td>12.38</td>
<td>12.38</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.00</td>
<td>0.00</td>
<td>12.46</td>
<td>12.46</td>
<td>12.46</td>
<td>12.46</td>
<td>12.46</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0.00</td>
<td>0.00</td>
<td>12.38</td>
<td>12.38</td>
<td>12.38</td>
<td>12.38</td>
<td>12.38</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td>72.45</td>
<td>72.45</td>
<td>72.45</td>
<td>72.45</td>
<td>72.45</td>
<td>72.45</td>
</tr>
</tbody>
</table>

(Continued...)

---

### Example 2

**Part 2 Exam Analysis**

**Correlations and means**

<table>
<thead>
<tr>
<th>Question</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>Correct</th>
<th>Item-</th>
<th>Correct</th>
<th>Item-</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>2</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>3</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>4</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>5</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Mean score</td>
<td>15.16</td>
<td>Standard deviation = 2.27</td>
<td>Cronbach's alpha = 0.92</td>
<td>0.052</td>
<td>Std. error of measurement = 1.186</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
who do well on that item (that is, answer it correctly) also tend to do
well on the entire test. The higher the positive correlation the
stronger this tendency.

Items with a positive correlation make a positive contribution
to the test’s consistency. Items with low item test correlations are
weak items; items with negative correlations are counterproductive
in that they reduce the consistency of the test. A negative correlation
means that persons who obtain high scores on the test tend to
answer that item incorrectly.

Since situations vary, there is no absolute value for an accept­
able item test correlation. As a rule of thumb, however, correlations
of at least 0.30 are desirable. If the test is to be refined for future
use, items with low or negative correlations should be replaced or
rewritten.

Mean

The mean score is the average test score obtained by the
respondents. It is the raw unweighted sum of correct responses
divided by the number of respondents.

Standard deviation

The standard deviation is an indication of the range of scores
in the test. If the distribution is normal the range will be three
standard deviations on each side of the mean. For example, in a
test with a mean of 70 and a standard deviation of 5, the highest
score will be about 85 and the lowest score about 55. Generally, the
higher the standard deviation the more widely scattered the scores
of the individuals taking the test.

Internal consistency

Internal consistency is measured with coefficient alpha when
the test scores are continuous and with Kuder Richardson 20 (KR20)
when the scores are dichotomous. When the scoring is
dichotomous (true or false, yes or no) coefficient alpha is identical
to KR20. Values range from negative (problematic) through 0 (heterogeneity)
range of +1.0 (perfect homogeneity). High positive values indicate that the items in
the test form a relatively homogeneous set, that is, that the individual
items are measuring something in common.

A low coefficient is a warning that the test contains items that
are only loosely related and that the test is impure since it appears to
be measuring several different attributes. In such a circumstance
the meaning of the test score is obviously obscure. Negative values,
which might occur if some items are inversely correlated, are
avoided by reverse scoring of discrepant items. What constitutes an
acceptable level of consistency varies with testing situation.

The standard error of measurement (SE) is estimated by.

\[ SE = \sqrt{T - \rho_{xx}} \]

where \( T \) is the test standard deviation.

Assuming that random errors of measurement are normally
distributed, the 95% confidence interval for \( T \) is obtained by the
formula,

\[ (T - 1.96 SE) \leq T \leq (T + 1.96 SE) \]

where SE is the standard error of measurement.

That is, there is a .95 probability that the true score lies within
1.96 standard errors of the unbiased estimate. Thus if a student
achieves a score of 82 and the test has a mean of 70, a standard
deviation of 6, and an internal consistency coefficient of .56, the student's true score is expected to fall in the range 71.2 to 86.8.

The student's true score is expected to fall in the range 71.2 to 86.8.

STUDENT RESULTS

Weights control the contribution that each question makes to
the test results. The weights and correct responses are shown
under a ruler designed to indicate question numbers from 1 to 100.
A blank for a correct response signifies an omitted question (Ex­
ample 3).

Rank shows the position of each student in the entire class
according to the number of correct responses. Low ranks signify
superior achievements. Ties are expressed with the highest of the
corresponding rank.

FINAL CLASS RESULTS

The total weights is the sum of weights for questions chosen on
the master sheet. Student names are listed in alphabetical order
with their ranks. Results are shown both for unweighted and
weighted scores. Each class or section begins on a fresh page
(Example 4).

The number of questions refers to results without weights.
Entries under weighted scores show the results when weights are
calculated.

Resultant percentages use weighted scores based on the total
weights of questions for which ovals were darkened on the master
sheet.

Guessing

The traditional view is that on multiple choice tests a correct
response might occur in two ways: (a) the respondent knew the
answer or worked it out logically, or (b) the respondent happened
to guess correctly. Since only knowledge is usually considered
relevant, guessing contributes to error variance. A guessing correc­
tion is possible under the assumption that the respondent either
knows an answer or selects one of the alternatives at random.
If there are \( k \) alternatives, the probability of a correct guess is \( 1/k \) and
the probability of an incorrect guess is \( (k-1)/k \). The odds are that a
subject can choose one answer correctly by chance for every \( (k-1)/k \)
wrong answers. An appropriate guessing correction is given in the
formula:

\[ R = \frac{C - W}{k - 1} \]

where \( R \) is the adjusted score (reflecting the subject's true
knowledge after accounting for guessing), \( C \) is the number of
correct answers and \( W \) is the number of wrong answers.
Example 3

Example 4
Unfortunately, empirical studies indicate that subjects rarely act in accord with the random guessing model and that psychometric gains resulting from the use of guessing corrections are minimal (Diamond & Evans, 1973). The testing system described here offers the following four choices, as shown in the table below.

<table>
<thead>
<tr>
<th>Points deducted from weighted correct answers</th>
<th>C</th>
<th>C-W</th>
<th>C-W/2</th>
<th>C-W/3</th>
<th>C-W/4</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All wrong</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Half of wrong</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>One-third of wrong</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>One-quarter of wrong</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**HARDWARE and SOFTWARE CONFIGURATIONS**

Data are collected on customized optical scanner sheets and transferred by an optical device to a file residing on a personal computer hard disk. A SAS/CONNECT session, activated by an operator, uploads the data to a simultaneous CMS session where the SAS 6.06 software described above is automatically invoked.

The chief administrative challenge was a conflict between integrity of student registration and efficiency of test administration. The registration process required that only students in good standing should have the right to answer an examination, and to that end, all answer sheets were preprinted with student names and identification numbers to prevent unregistered students from gaining academic standing. Missing or erroneous examination forms could not be replaced or corrected by hand at the last minute because forms were preprinted with black marks ("slugs") that did not correspond to ovals printed on standard forms. A new custom-designed form that accommodated machine slugs and manual entry was implemented in 1990.

Fields for age and sex were added to the new form to permit its use in attitudinal and public opinion studies.

**REFERENCES**


